algorithms

MDPI

Article

Reddit CrosspostNet—Studying Reddit Communities with Large-Scale Crosspost Graph Networks

Jan Sawicki [1,2,*,†], Maria Ganzha [1,*,†], Marcin Paprzycki [3] and Yutaka Watanobe [2]

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland
[2] Department of Computer Science and Engineering, The University of Aizu, Aizu-Wakamatsu 965-8580, Japan; yutaka@u-aizu.ac.jp
[3] Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland; paprzyck@ibspan.waw.pl
* Correspondence: jan.sawicki2.dokt@pw.edu.pl (J.S.); maria.ganzha@pw.edu.pl (M.G.)
† These authors contributed equally to this work.

Abstract: As the largest open social medium on the Internet, Reddit is widely studied in the scientific literature. Due to its structured form and division into topical subfora (subreddits), conducted research often concerns connections and interactions between users and/or whole, subreddit-structure-based communities. Overall, the relations between communities are most often studied by applying graph networks, with various creation algorithms. In this work, a novel approach is proposed to build and understand the structure of Reddit. It is based on crossposts—posts that appeared on one subreddit and then were crossposted to another. After capturing one year of crossposts, a directed weighted graph network, using seven million posts from over 10,000 of the most popular subreddits, has been created. Using graph network algorithms, its characteristics are captured and compared to similar studies. We identify the information "sinks" and "sources"—the most active crossposting subreddits. Moreover, we obtained graph network metrics: the degree (modeled with the Power Law), clustering, community detection algorithms, and connected components structure network are compared to previous studies on Reddit network(s), yielding consistent, but also novel results. Finally, the relations between extensively studied subreddits (e.g., r/AITA, r/Parenting, r/politics) and new ones, which were not accounted for in previous research, opening new paths for data-driven studies, are summarized.

Keywords: graph network-based analysis; Reddit; subreddits; online social networks; big data; crossposts

Citation: Sawicki, J.; Ganzha, M.; Paprzycki, M.; Watanobe, Y. Reddit CrosspostNet—Studying Reddit Communities with Large-Scale Crosspost Graph Networks. *Algorithms* **2023**, *16*, 424. https://doi.org/10.3390/a16090424

Academic Editor: Adele Anna Rescigno

Received: 31 July 2023
Revised: 23 August 2023
Accepted: 28 August 2023
Published: 4 September 2023

1. Introduction

Reddit is the largest public, topically separated social network, on the Internet. Its unique structure, consisting of thematic subfora (subreddits), has been widely studied. Most often, the subreddit network has been modeled with graph networks capturing user interactions, posts, comments, and authorship. However, an important Reddit feature is underutilized in this context. This is the existence of crossposts, which represent subreddit-to-subreddit relations. Crossposts are posts that have been copied (crossposted from one subreddit to another while preserving the original source post). A crosspost always has a direction, i.e., it goes from a subreddit to a subreddit. Obviously, the existence of a given crosspost indicates that the users of one of the subreddits believed that this post would be of interest to the readers of another subreddit. Following up this line of reasoning, note that crossposts have a "direction" with a "source" and a "sink". Moreover, the number of crossposts between two subreddits can be seen as a measure of the strength of a relationship. In other words, if two subreddits are connected by a very large number of crossposts, it indicates that their readers believe that there is a strong, shared interest. Therefore, directed weighted (by post count) graph network(s) can be built on the basis of existing crossposts.

In this work, an algorithm is used to automatically build a large-scale directed weighted graph network, on the basis of 7,643,447 crossposts that appeared between the top 131,477 subreddits (understood as subreddits with the largest number of users). Next, the resulting graph network is analyzed. In particular, various network metrics are calculated and compared to previous results obtained when studying Reddit's graph structure. In addition to macro-scale observations, individual relations between subreddits are studied, to capture those that receive or produce most crossposts and to find relations that have not been previously accounted for in Reddit-related research. These observations are turned into a proposed extension to the existing approaches, used in topical studies based on Reddit. This extension to the standard data pipeline may help future studies by suggesting additional subreddits of interest.

Please be advised that this work contains explicit names of the subreddits with pornographic and vulgar themes, which have not been censored to present the full analysis without sections.

2. State-of-the-Art of Subreddit Graph Modelling

Reddit and other social networks are extensively studied with graph networks. The most common approach is based on capturing user-to-user relations. Here, connections (in the graph) are built using, for instance: (i) user co-commenting on post(s) [1], (ii) users commenting on each other posts [2], or (iii) a "question-person & answer-person" model [3] (where people who ask questions are conceptually connected with those who respond to them).

Specifically, the study [1] provides a specific model for forming connections between users, based on their co-commenting. They account not only for the co-commenting, but also for the number of commented posts and the similarity of comments. Here, the weight of the connection is a product of the number of comments and the similarity measure. The latter one is called "mutual interest" and it is defined as $(P_{si} \cap P_{sj})/(P_{si} \cup P_{sj})$, where P_{si} denotes the number of posts the user U_{si} commented on and P_{sj} denotes the number of posts the user U_{sj} commented on. The authors use a dataset of 54,504,410 comments with 2,611,445 authors. They take three samples and a frequent set of users and build three networks. All of them contain about 9700 nodes and about 670,000 edges (after data cleaning). Next, the study analyzes the basic social network properties of said graphs. Moreover, community detection is performed and its results are analyzed. The key results of this work will be further discussed in the context of the network built in this contribution.

In the second study [2], the authors gather data from 24,298 subreddits, posts (a.k.a. submissions), and comments. Then the inter-subreddit relations are modeled, based on links in the subreddit description, resulting in a graph with 27,091 nodes and 9050 edges. The completed analysis includes measuring the distribution of node degree, triangles, largest connected component, and correlations between network measures and subreddit subscribers count. Furthermore, a tree-like structure of comments of posts is analyzed. Here, comments form a tree graph—a directed acyclic graph and the conducted research checks how many nodes (subreddits) have what number of comments and what is the tree depth. Furthermore, three new graphs are built. They are based on users, with 31,731 nodes, 1,804,974 edges (the "loose graph"), 176,385 nodes and 435,413 edges (the "tight graph"), and 13,802 nodes and 13,251 edges (the "strict graph"). The edges in the loose graph exist if user A commented on user B (undirected). The edges in the tight graph exist if user A commented on user B and user B commented on user A. The edges in the strict graph exist if user A commented on User B four times and vice versa. For these three graphs, the study measures (again) the basic network statistics, such as the node degree and the connected components. The results of the study introduce the analyzed dataset and summarize the findings, though without additional interpretation.

The last study in this group (ref. [3]), analyzes the so-called "question-answer" model. Here, only the "Ask" subreddits, AskScienceDiscussion, AskMen, AskScience, AskWomen, CompSci, DesMoines, IAmA, MachineLearning, Movies, MyLittlePony, PersonalFinance,

TalesFromTechSupport, and WashingtonDC, subreddits are studied. This is because the posts in these subreddits are mostly questions, while the comments provide answers. Using the top 100 posts from one month and an average of 200 comments per post, a network is instantiated. The nodes are users and the directed edges are defined if an "answering" user commented on posts of a "questioning" user. This way, the question users are nodes with very high in-degree and low out-degree, while the answering users are the opposite. Although this analysis focuses on a very particular relation, it uses very universal network analysis methods. These are in-degree and out-degree analysis, clustering coefficient, triangle density, and also subscriber count analysis of subreddits. The conclusion is that identifying question-answer roles on Reddit is clearly possible.

Overall, these studies utilize multiple concepts that can be used when analyzing social graph networks. Among them, the most popular are the network metrics (e.g., clustering, average shortest path, diameter, components) and distribution metrics (e.g., degree, degree centrality). However, it should be noted that the focus of those works is on user-user interactions. As such, it is missing knowledge that can be gained on different levels of granularity.

Another research direction dealt with microscale community-specific studies, which model community interactions. However, typically only a fraction of Reddit was studied. They model relations between particular communities, e.g., gender-related subreddits [4], cybermarkets [5], politics [6,7], mental health [8], or the "Ask" subreddits [9]. These studies are conceptually close to the analyses presented in what follows, because they deal with "Reddit communities". However, the scope of their expertise is beyond computer science and they focus more on social, medical, and other aspects. Moreover, these contributions focus only on microscale topic-specific subreddits and their networks, while this work models an almost complete spectrum of subreddits.

Finally, a recent study (ref. [10]) approaches the problem of graph-based modelling of the complete subreddit structure of Reddit. It uses "subreddit link mentions" to characterize subreddit connections, which are then analyzed using different graph metrics. The link mention appears when a user mentions a different subreddit, either in the post title, the content, or in the comments. The link does not have to be a full link, e.g., https://reddit.com/r/Science (accessed on 22 August 2023), it can also be "r/Science". Importantly for this contribution, that study [10] suggests the need to use and study crossposts in future works. This can be seen as an additional justification for building and analyzing Reddit CrosspostNet, a graph network, which is modeling intersubreddit relations on the basis of the existence and multitude of crossposts.

3. Dataset and Network Creation

As noted, the aim of this work is to propose a new approach to analyzing the structure of and relations between, subreddits using graph network algorithms applied to represent the crosspost perspective. In this context, we first introduce the dataset used in the reported work. In the first step, all crossposts, from all subreddits with over 10,000 subscribers, have been collected for the period 1 January 2022–31 December 2022. The crossposts have been then filtered to remove insignificant content. Specifically, crossposts with a score (Reddit's "appreciation" mechanism [11]) of less than 2 were removed. Note that, by default, all posts obtain a score of 1 when posted. Therefore, it was assumed that posts with "no appreciation" can be seen as "inconsequential" from the perspective of Reddit as a medium of information exchange. For similar reasons, all posts with 0 comments were also removed. Again, it was assumed that "nobody cares" about posts when they have no comments. Obviously, these values are arbitrary. It could have been decided to keep crossposts with more user involvement (e.g., a minimum score of 5 and a minimum of 5 comments). In this way, only the "most influential" crossposts could be kept and used in further analysis. However, it has been assumed that, in this contribution, the focus will be on a complete picture of the information structure of Reddit. Studying the effects of including only the most influential crossposts is left to future work. Finally, all crossposts with the topic "[deleted]"

or "[removed]" were also pruned, as these were either deleted by the author or removed by Reddit's moderation. Completion of the pruning process resulted in a dataset consisting of 7,643,447 crossposts from 131,477 subreddits. Here, due to the recent restrictions imposed by Reddit, we have decided to not post the dataset publicly online. However, it may be provided to academic researchers (only), upon contacting the lead author.

In the second step, a graph network was formed, with subreddits as nodes and crossposts capturing the directional relationship between them. The edges were weighted by the number of crossposts. The pseudocode of the graph network building algorithm is as follows: in the Algorithm 1.

Algorithm 1 The algorithm of constructing the CrosspostNet

crossposts ← all available crossposts after cleaning (a crosspost is a dictionary with values "from subreddit" and "to subreddit")
$V \leftarrow \{\}$ (empty set of vertices; it will become a set of values)
$E \leftarrow \{\}$ (empty set of edges; it will become a set of ordered pairs (node1, node2) with an additional value indicating weight)
while $|crossposts| > 0$ **do**
    test
    $crossposts \leftarrow crossposts[0]$ (take the first element)
    $i \leftarrow i - 1$
    if $crossposts["from\ subreddit"] \notin V$ **then**
        $V \leftarrow V + crosspost["from\ subreddit"]$
    end if
    if $crossposts["to\ subreddit"] \notin V$ **then**
        $V \leftarrow V + crosspost["from\ subreddit"]$
    end if
    $new\_edge \leftarrow (crossposts["from\ subreddit"], crossposts["to\ subreddit"])$
    if $new\_edge \notin E$ **then**
        $E \leftarrow E + new\_edge$
    end if
    $new\_edge\_weight \leftarrow new\_edge\ weight + 1$
    $crossposts \leftarrow crossposts \backslash crosspost$
end while

As a result, the set V becomes the set of vertices (labeled with subreddit names) and the set E is the set of directed edges, weighted by the number of crossposts between the subreddits.

For example, there were 2083 crossposts from r/Unexpected to r/HolUp and 33 crossposts from r/HolUp to r/Unexpected. So, the node representing r/Unexpected is connected to r/HolUp with an edge of weight 2083, while r/HolUp is connected to r/Unexpected with an edge of weight 33.

The final graph has 92,505 nodes (subreddits); connected by 697,355 directed edges and was created in 6.97 s using an AMD EPYC 7742 64-Core Processor with networkx library 2.6.3 in Python 3.9.

Note that the runtime measured only the graph creation based on the already preprocessed data. In what follows, the network will be named Reddit *CrosspostNet*. Obviously, the obtained graph network is too large to comprehensively visualize. However, with appropriate filtering, it is displayed in an interactive form in the data repository of this project at https://github.com/JanSawicki/CrosspostNet (accessed on 22 August 2023) The presented network contains only edges (connections), which had the minimum weight of 10, meaning that there were at least 10 crossposts between 2 nodes (subreddits) and had at least one neighbor (degree > 0) and the summed score of them was at least 2. An illustrative screenshot of the graph is presented in Figure 1. Moreover, Figure 2 presents an illustrative fragment of the network, zoomed-in on the nodes with the highest in- and out-degree.

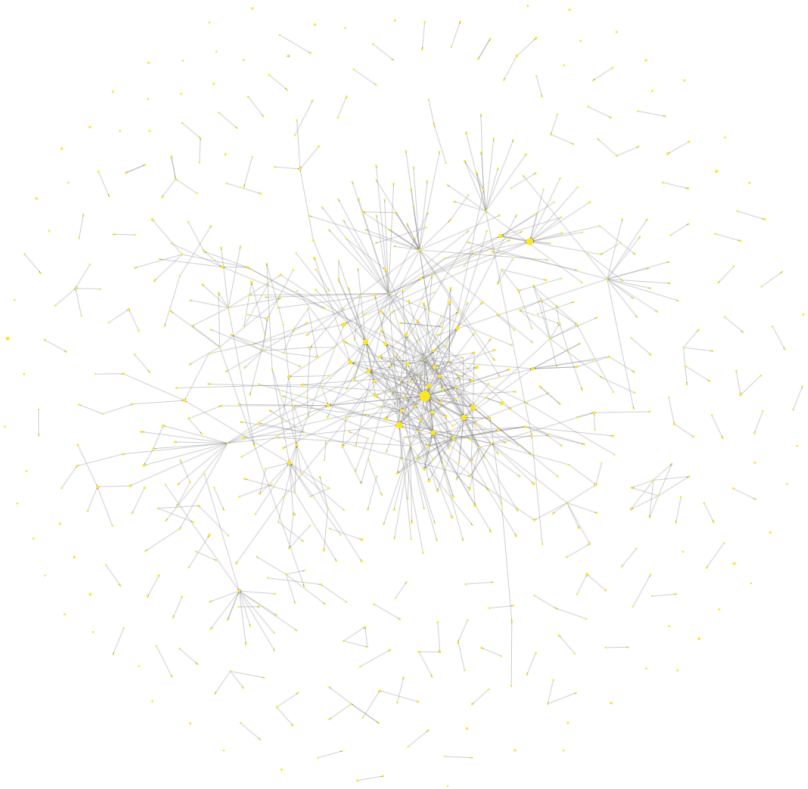

Figure 1. CrosspostNet illustrative visualization (simplified).

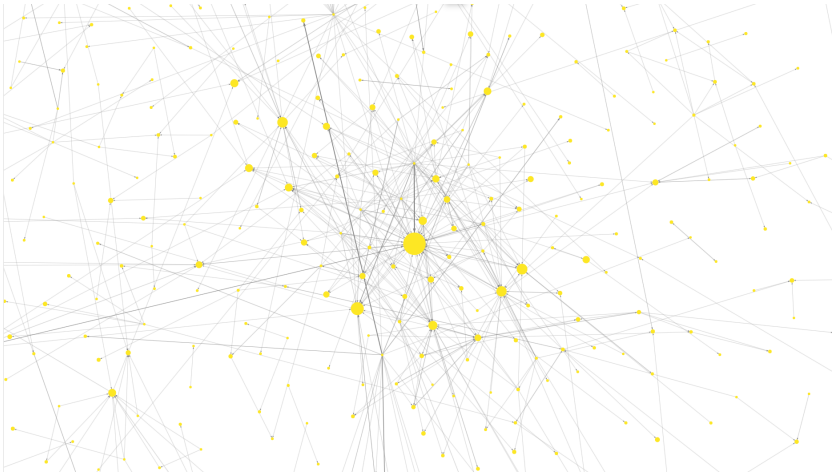

Figure 2. CrosspostNet illustrative visualization (simplified and zoomed).

4. Characterizing CrosspostNet

Having built the crosspost-driven subreddit structure representing the graph network (CrosspostNet), we can move on to its characterization and a comprehensive description of its key characteristics. In this section, the macro-scale metrics of the graph network are provided and compared with the most important and recent studies modeling Reddit structures(s) using graphs.

4.1. Node Degree Analysis

We start the analysis of CrosspostNet by discussing the node degrees. Figures 3 and 4 depict the top nodes, sorted on the basis of their in- and out-degrees, respectively.

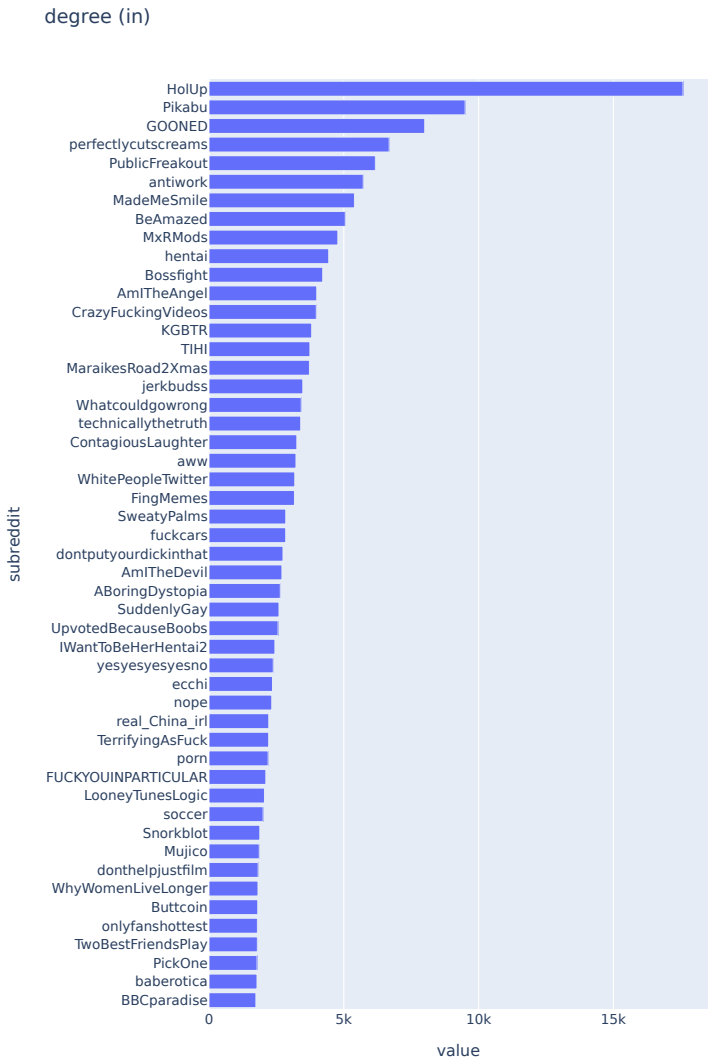

Figure 3. Top 50 subreddits (nodes) by in-degree.

The first thing to notice is the Power Law nature of both distributions. This is a common phenomenon in social networks. The exponent of the degree distribution of a scale-free network (γ) of the distribution usually lies between 2 and 3 [12]. As established in similar studies of Reddit social networks [1], the Power Law should materialize also here. After simple calculations, it was found that for the CrosspostNet, the resulting values for the degree distribution (Figures 3 and 4) are $\gamma = 2.83$ (in-degree) and $\gamma = 2.35$ (out-degree). Therefore, as expected, CrosspostNet is a scale-free network (see, also [12]).

We now focus attention on the top in- and out-crossposting subreddits. It is clearly visible that r/HolUp (acronym from "hold up!"—a phrase used on the Internet when something unexpected happens) is the major subreddit, to which crossposts are posted. It is the main receiver of crossposts in the whole of Reddit. On the other hand, r/interestingasfuck, r/Unexpected and r/damnthatsinteresting produce most posts, which are then crossposted to other subreddits. They are the main senders (producers) on the platform. Interestingly, R/HolUp and r/Unexpected consist of mostly pictures or GIFs that show a situation with an unusual "twist" (the so-called "holup!" moment). R/interestingasfuck and r/damnthatsinteresting, as their names suggest, display pictures that are interesting to users, mostly trivia, unintuitive facts, and other random pieces of knowledge. Note that all these subreddits do not focus on a particular topic, but rather a feeling that the user

experiences when browsing them. This means that there is no particular topic that attracts sharing and crossposting, rather a general and/or unspecified content.

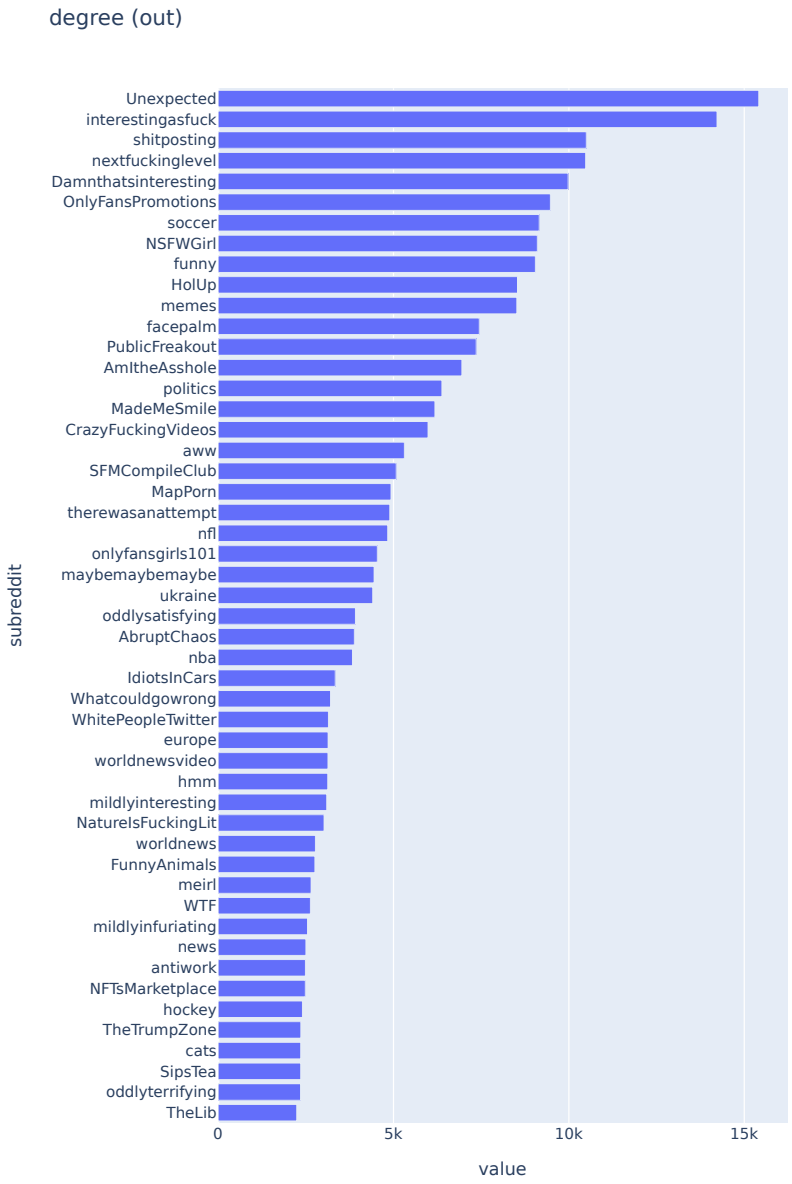

Figure 4. Top 50 subreddits (nodes) by out-degree.

In the following analysis, the "in-" and "out-" prefixes refer to the direction of the edges (hence the degree). The in-degree refers to the in-going edges, which model the incoming crossposts (posts that were crossposted to the subreddit). The out-degree refers to the out-going edges, which model the out-going crossposts (posts that were crossposted from one subreddit to another). The top subreddits in both rankings (the in-degree and the out-degree) consist of subreddits with very broad topics, such as r/nextfuckinglevel (pictures and videos of skillfully performed activities), r/funny (funny pictures and videos), etc. However, there are several subreddits that are more narrow topic-wise. At the top of subreddits with the highest in-degree (subreddits that receive the most crossposts), there is r/pikabu, a Russian subreddit which consists of various content in Russian and r/antiwork, a subreddit about getting "the most out of a work-free life". In the top of subreddits with the highest out-degree (subreddits that produce posts that are then crossposted) there is also r/soccer.

We now consider the average neighbor degree. It is a standard graph metric, which captures the average of degrees over the set of all nodes that are neighbors of a given node. Hence, it indirectly captures how well-connected a node is. The four cases (related to the fact that CrosspostNet is a directed graph), of average neighbor degree, are represented in: Figures 5 and 6 show top nodes by average out-neighbor in- and out-degree. Figures 7 and 8 show top nodes by average out-neighbor in- and out-degree.

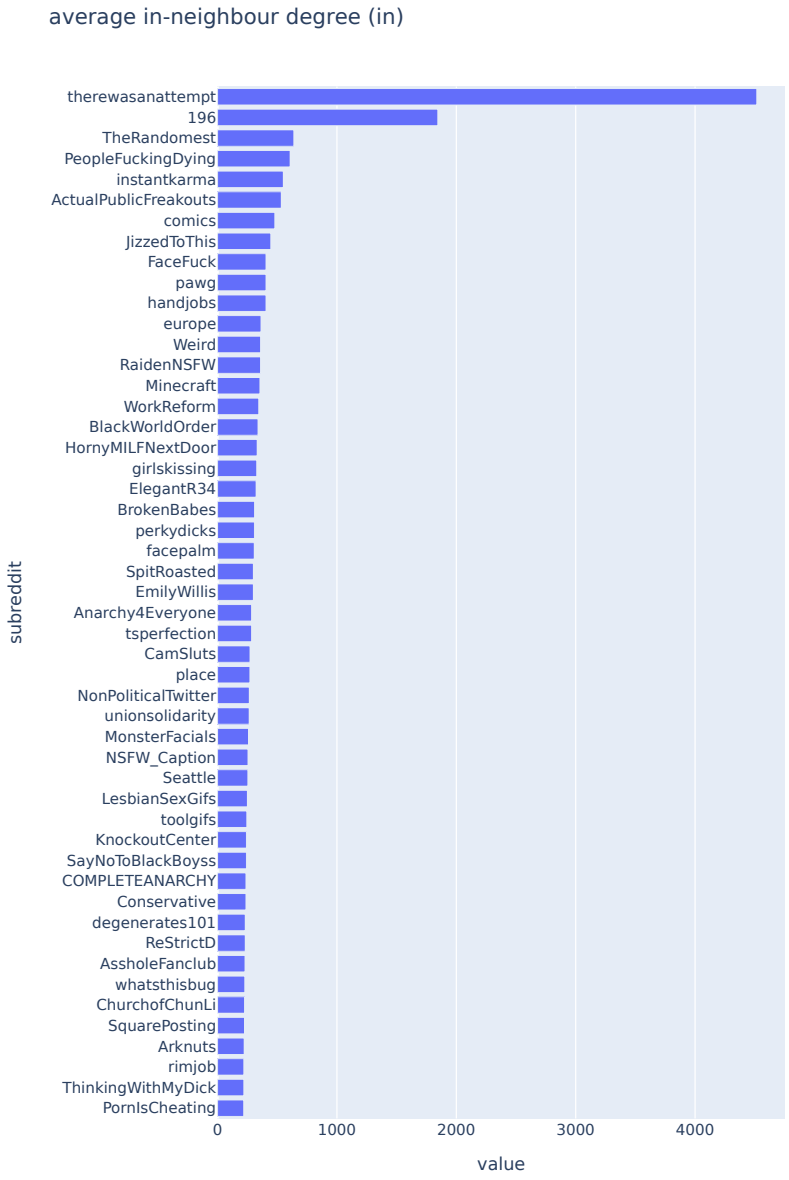

Figure 5. Top 50 subreddits (nodes) by average in-neighbor in-degree.

First, we notice that while the situation with a power-like distribution is similar for the average in- or out-neighbor in-degree, the out-degree shows a nearly uniform distribution for both in- and out-neighbors. The Power Law coefficients are as follows: in-neighbor degree (in-): 3.44; out-neighbor degree (in-): 2.93; in-neighbor degree (out-): 2.11; out-neighbor degree (out-): 2.66. Even though the number may seem similar, the distribution of in-degrees is much different than the out-degrees. This shows that while there is a big disproportion between nodes' neighbors in-degree (some nodes attract crossposts from "multiple directions"—they are very popular to be crossposted in), the neighbors out-degree is very similar for all nodes (it is not possible to point to a group of subreddits that are main sources of "propagating information"). This translates to, what we believe to be an interesting finding. Most subreddits receive crossposts with subreddits, which rarely

have crossposts themselves. On the other hand, practically all subreddits exchange (in- and out-degree) crossposts with subreddits of similar out-going crossposts (out-degree). While the uniform distribution of out-degrees does not indicate any outstanding subreddits, the power-like distributions are worth taking a closer look at. The top of both in- and out-neighbors contains the following subreddits: r/therewasanattempt, r/196 and r/instantkarma, r/TheRandomest, r/PeopleFuckingDying. All of them contain various types of memes in an image, or video format that have no particular topic. The content of r/therewasanattempt is finishing the titular sentence with a (usually funny) comment about a given image or video. R/196 has recently become restricted and contains all types of meme-like images. R/instantkarma contains content that shows when someone, or something, was instantly punished for something that they did. R/TheRandomest is described as: "The weirdest place on Reddit. Some come for giggles, some come for awe, some come for shock and some come for anything in between." R/PeopleFuckingDying is not about human death, but about hyperbolized reaction to some event, usually posted as an image or video.

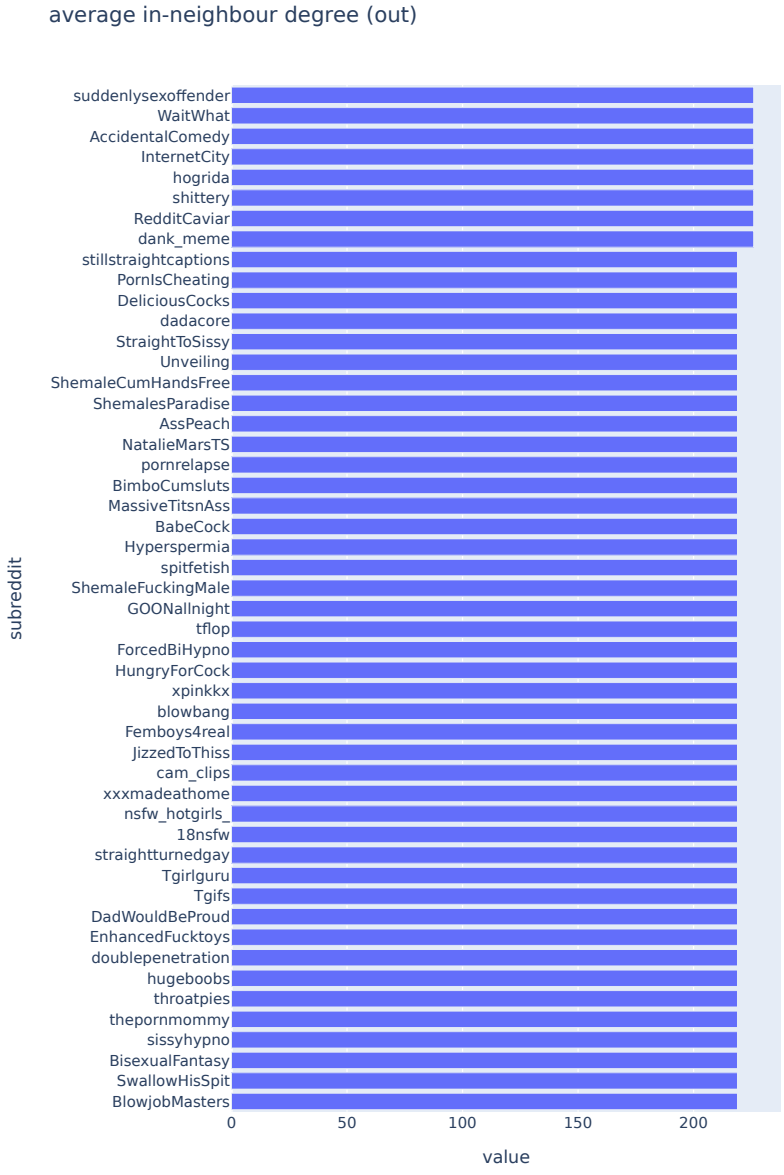

Figure 6. Top 50 subreddits (nodes) by average in-neighbor out-degree.

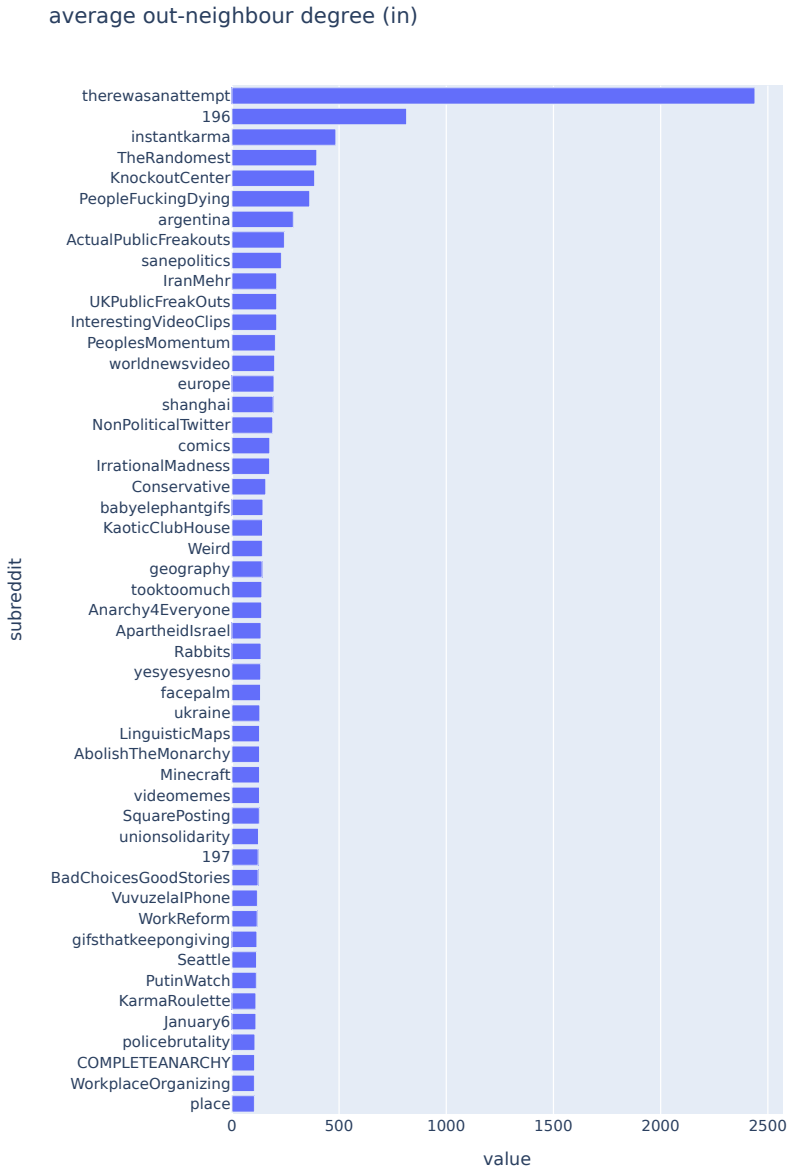

Figure 7. Top 50 subreddits (nodes) by average out-neighbor in-degree.

We note that the results presented above are very similar to those reported in other studies, for instance in [1,2] that have been modeling Reddit content and users. Those studies also found a power-like distribution of node degree. This led to the conclusion that Reddit is subject to a "small-worlds" phenomenon, where most nodes (subreddits) are connected with few steps, because of the existence of hubs (nodes with very high degrees). Moreover, the power-like distribution was also present there and explicitly called a scale-free network.

While the previous studies were focused on users, this contribution shows that the same applies to community-community interactions. Specifically, also in this case, the small-worlds and the scale-free phenomena materialize here. Practically, this shows that the content flow (both in- and out-crossposts) usually is pushed into or comes from a large hub-like group of subreddits. Interestingly, this seems to represent a well-known stage of random graph evolution [13] where the degree distribution is also consistent with the Power Law distribution.

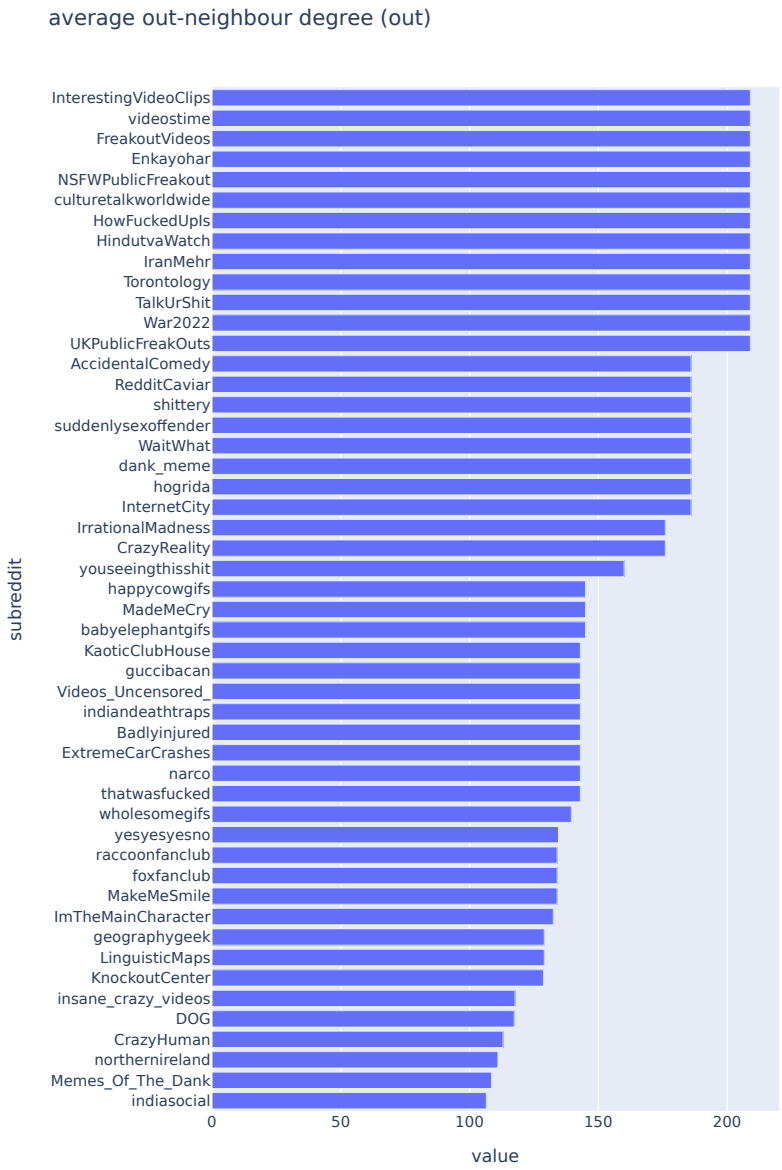

Figure 8. Top 50 subreddits (nodes) by average out-neighbor out-degree.

4.2. Edge Analysis

We now take a look at the edges existing in CrosspostNet. Table 1 presents top-directed edges and the subreddits they join.

There are several interesting observations. First, the largest numbers of crossposts are between three subreddits: r/AmITheAngel, r/AmITheDevil, and r/AmITheAsshole. All these three communities are used to ask questions and seek the moral judgement of other users, who vote as to whether the behavior of the poster was good ("Angel"), bad ("Devil", or "Asshole"). Interestingly, these social behaviors and the r/AmITheAsshole subreddi have a dedicated study and this finding is further elaborated in Section 5.2.

Next, as expected, there are r/HolUp and r/Unexpected. These two communities already appear at the top of rankings, based on the node degree (see, above). What is interesting is that the relationship between r/HolUp and r/Unexpected (but also r/shitposting and r/memes—both communities about Internet memes) is very much one way. All three of those subreddits crosspost most often to r/HolUp, but there is not an edge in the top of edges from r/HolUp to any of these, or any other, subreddits. As of what concerns the rest of the ranking contains mostly pornography-oriented subreddits, a.o. r/NSFWGirl, r/babeerotica, r/BabesNSFW, etc.

Table 1. Top subreddit connections (edges) by crossposts count (edge weight).

To Subreddit	From Subreddit	Crosspost Count
AmITheAngel	AmItheAsshole	3461
AmITheDevil	AmItheAsshole	2504
HolUp	Unexpected	2083
baberotica	NSFWGirl	1754
BabesNSFW	NSFWGirl	1334
onlyfanshottest	OnlyFansPromotions	972
HolUp	shitposting	970
HolUp	memes	969
Slutsofonlyfans	OnlyFansPromotions	947
OnlyfansXXX	OnlyFansPromotions	919
onlyfanschicks	OnlyFansPromotions	909
GoneWildOnlyfans	OnlyFansPromotions	899
AdorableOnlyfans	OnlyFansPromotions	894
NaughtyOnlyfans	OnlyFansPromotions	893
HotOnlyfans	OnlyFansPromotions	891
reddevils	soccer	879
OnlyfansAmateurs	OnlyFansPromotions	878
WhitePeopleTwitter	TheLib	877
vexillologycirclejerk	vexillology	850
Wives_NSFW	NSFWGirl	819
Tenshigao	NSFWGirl	792

However, there is one group of subreddits that is particularly outstanding. It is even visible to the naked eye when performing the graph network visualization. These are the subreddits related to the OnlyFans platform. OnlyFans is "the 18+ subscription platform, empowering creators to own their full potential, monetize their content and develop authentic connections with their fans" (https://onlyfans.com/about, accessed on 22 August 2023). This group contains subreddits: r/OnlyFansPromotions, r/Slutsofonlyfans, r/OnlyfansXXX, r/onlyfanschicks, r/Gone/-Wild/-Onlyfans, r/AdorableOnlyfans, r/NaughtyOnlyfans, r/HotOnlyfans, r/Only/-fans/-Amateurs. These subreddits are not thematically related to the OnlyFans platform itself, but rather they have been created for the promotion of content creators, on said platform. The r/OnlyFansPromotions subreddit is central, as all other subreddits crosspost to it, while there are no known crossposts back to any of them. Specifically, there is not a single crosspost in the dataset from r/OnlyFansPromotions to another subreddit. Obviously, there could be some such subreddits in the times outside of these represented in the dataset, but there is no reason to believe that the existing dataset is an exception. A manual exploration showed that practically all crossposts into r/OnlyFansPromotions contain marketing content, promoting OnlyFans creators, which is confirmed by the name and purpose of this central subreddit. This shows that crossposts are used not only as a cross-reference between communities of similar interests. They can also be used for self-promotion. The fact that most crossposts are related to an 18+ website somewhat resonates with the fact that the biggest cluster of subreddits (identified when clustering was applied to them) is related to pornography (see, above and [14]).

The last two observations are that r/soccer frequently crossposts to r/reddevils (a Manchester United soccer club subreddit) and r/vexillology crossposts often to r/vexillolo/-gycirclejerk (both subreddit related to the study of flags). Interestingly, these subreddit pairs crossposts both ways. R/reddevils has crossposts 256 (99th percentile) to r/soccer meaning that although the crosspost relation is not exactly balanced, it is mutual. However, r/vexillologycirclejerk has only 10 crossposts (96th percentile) to r/vexillology. This shows an asymmetric relation between those of the flag-related subreddits.

4.3. Paths Sutructure

Another important aspect of a graph network is the graph structure, including the internode distances. The original work [10], measured the *average shortest path* between

nodes, defined as the average of all shortest paths between all node pairs, and the diameter defined as the maximum shortest path between nodes. An important issue in finding the shortest path between nodes is that a path needs to exist at all. Therefore, before calculating traversability metrics, we have to account for connectivity and check what nodes have a path between them in both directions. The first finding was that the CrosspostNet is not strongly connected, meaning that not all vertices have paths between them. Therefore, calculating the shortest path and the diameter would return an infinity, as some of the nodes have infinitely long paths (i.e., no path at all). Hence, the largest connected component, which is the largest subset of nodes that is connected with paths, had to be found. It has been established that for the subreddit-based crosspost driven graph network, the largest component consists of 7660 nodes (subreddits). This constitutes 78% of all nodes in the graph. However, the total number of components found in the graph network is 796. As expected, the remaining components are extremely small. The second-largest component contains 25 subreddits (3.1%). Over 693 (87%) of the components contain a maximum of 3 crosspost connected subreddits. This shows that, while the main world of inter-subreddit relations coalesces into one large connected component, there are many micro-worlds where discussions and crossposts interaction happen. This again resonates with the previously mentioned small-world phenomena.

Having selected the main component (with 7660 subreddits), we now apply the algorithm for finding the shortest path (the Dijkstra's [15] algorithm) and the CrosspostNet diameter (a measure based on the eccentricity algorithm, as shown in [16]).

Please note that this is a directed graph, so according to the definition of the average shortest path, there are two possible connections between each node (both directions), and if there is no directed path between two nodes, their shortest path length is 0 (not infinity, for practical reasons). Analysis of the largest connected component allowed establishing that only 4.2% of all node pairs have a directed path between them. This is an important observation that the crossposting relations are in most cases one way, which prevents a long-distance flow of information, while causing strong short-distance relations between subreddits. Therefore, calculating the shortest path does not make much sense, as it would be extremely small (due to zeros for unconnected nodes) and would not model the actual distances.

As a side note, it is also worth mentioning that in the whole population of crossposts, there have been just four (less than 10e−7%) examples of crossposts, which were crossposted more than once. This means that most crossposts appear once (originally), when they are crossposted and the original posts are extremely unlikely to be crossposted again. This and the previous observation indicate that crosspost-based interactions are extremely localized and this should be kept in mind when considering results reported in this contribution.

As a result of the above-mentioned findings, it has been decided to calculate the paths after converting the graph into an undirected one. The average shortest path in the resulting undirected graph (largest component) is 5.91, while the diameter is 16. This resonates with the results of the previous work [10], where the value of "95% effective" diameter was between 15 and 20, depending on the time. On the other hand, the distances between nodes of the link-based network [10] were growing in time from 3.2 to 3.6. CrosspostNet has about 1.5 times longer distances between the nodes. Nevertheless, taking into account the size of the graph, these results are clearly of the same order and can be seen as supporting each other.

Overall, such a short average distance between the nodes and the small diameter (with regard to the graph size) suggests that the CrosspostNet is, again, subject to small-worlds phenomena, where most of the nodes are connected indirectly via large-degree hubs. Having discovered that, it is reasonable to explore the "small worlds" (communities) with clustering and community algorithms.

4.4. Community Analysis

Here, first, we consider the average clustering coefficient in the undirected graph, which measures the average of local clustering coefficients (the fraction of triangles that exist divided by all possible triangles in the node's neighborhood). The obtained value is 0.153. It is similar to the one reported in the previous study (ref. [10]), where it oscillated in time between 0.1 and 0.2. This shows that approaching the Reddit network from the point of view of cross-linking and crossposting produces communities with a similar number of possible triangles in the graph.

Going further, the degree assortativity can be checked. This measure explains the correlation between node degree and its neighbor node degree. The degree assortativity (Pearson correlation coefficient) ranges between -1 (disassortative network, i.e., nodes connect to nodes with different degrees) to 1 (assortative network, i.e., nodes connect to nodes with similar degrees). In the previous study [10], the reported value was about 0.0. Here, the calculated value was -0.04. This means that the link-based graph, studied in 2020 and the crosspost-driven subreddit graph, from 2022, are both neither assortative nor disassortative. In other words, there is *no relationship* between the degree of a given node and that of its neighbors.

Next, we consider the community analysis. Studies on user-based social networks showed interesting results when applying algorithms for community discovery, such as the Louvain method [17,18]. The first work [17] analyzes inter-subreddit conflicts of 2016. It creates a co-conflict network, where nodes are subreddits, and applies different algorithms and methods to analyze it. Among them, the community detection is used to determine, which groups of subreddits (rather than pairs) are co-targeted. What is important is that the authors mention three community detection algorithms (FastGreedy [19], InfoMap [20] and Louvain methods [21]), briefly compare them, and, finally, decide that the Louvain algorithm is the best for the Reddit graph network. Louvain's method is based on modularity maximization (a measure of cohesiveness of the network). The authors chose the Louvain algorithm over the remaining algorithms because it "follows a hierarchical approach by first finding small, cohesive communities and then iteratively collapsing them in a hierarchical fashion (. . .) produces reasonably sized communities and the results of the community detection algorithm were very stable". Learning from the findings of that study, the Louvain method is also applied here, to find the communities of subreddits connected with crossposts.

The Louvain algorithm found 796 different communities. Among them, there are 575 (72%) communities of size 2 and 117 of size 3 (15%). This totals over 87% of all communities consisting of 2–3 subreddits. On the other hand, 4 large communities have been detected. The largest community contains, 1334 subreddits and there is no particular theme among them (as far as we could establish on the basis of manual inspection). The subreddits are about multiple topics, ranging from politics, to memes, to pictures of animals. This means that the majority of crossposts "connect" vaguely related subreddits, i.e., there is no particular largest thematic community of subreddits that has strong connections based on crossposts. However, the second-largest community (876 subreddits) has a very visible theme. All subreddits in the second-largest community contain mostly pornography. The third-largest community (571 subreddits) contains mostly subreddits that concern politics, or are related to individual countries. Finally, the last one (539 subreddits) is also about pornography, but with a definite focus on the drawn/animated works (with hentai drawings represented in the majority of them).

It should be noted that communities discovered using the Louvain method are quite similar to these reported in a recent study [14], where subreddit clusters were established based on their content, modeled with textual embeddings. In that study, the largest clusters also included pornography, politics, and memes, as the top largest communities. Therefore, it can be stated that results obtained on the basis of crosspost connections are consistent with these established for content similarity. Moreover, again, results obtained using two different approaches support the validity of each other.

Embedding Similarity

In addition to finding subreddits that are the biggest crosspost contributors or form node communities, it is possible to approach the graph network seeking subreddits that are similar in terms of their graph structure. While an existing edge is the most direct connection, it is also possible to look for subreddits, which have similar architectural positions in the whole network. Previous studies on social networks (e.g., [22]) discovered that node embeddings can retain a combination of network properties different from the standard metrics (such as degree, centrality, etc.). Therefore, the most prominent solution, mentioned in [22], Node2Vec [23] has been applied.

Node2Vec is an algorithm that allows embedding graph network nodes and building vectors representing them. The algorithm is based on random walks in the graph. Here, for each node in CrosspostNet, a vector representing the node is generated using the Node2Vec random walks strategy. Node2Vec has different hyperparameters, such as p—controlling how deep should the algorithm go, q—controlling the chance of immediately revisiting a node, vector size, number of walks, and walk length. In performed experiments, different hyperparameters for CrosspostNet, using combinations of p (1, 2), q (1, 2), vector size (16, 32, 64), walk length (10, 100, 200), number of walks (100, 200) have been explored. None of them yielded, results that were much different in terms of vector similarity, which is the next step. Hence, results with p = 2, q = 2, vector size = 64 and walk length = 100 are reported.

Having generated the vectors, the next step is to find the most similar ones to merge them into groups. Again, vector similarity can be measured with many algorithms. The cosine vector similarity was chosen, because it showed the best results in other vector similarity applications (e.g., [24]). The each-with-each similarity of vectors (representing subreddits) has been used. Next, the resulting pairs were sorted by similarity, and the top pairs were extracted. They are presented in Table 2. Note that Table 2 disregards nodes of degree less than 3 since their structural embedding is trivial and their similarity is obvious and not meaningful.

Table 2. Top subreddits (nodes) groups in the top 99.9th percentile cosine similarity of node embedding with Node2Vec.

Subreddit Group	Topic
r/Vilen_Collection, r/Desinude_daily, r/Desi_Kama, r/Lust_maal, r/indianEXnude, r/Fap_, r/DesiXvideo, r/premium_Daily, r/Desii, r/Lustdesii, r/Desi_Boners	Pornography
r/PremierLeague, r/LiverpoolFC, r/FantasyPL, r/Barca, r/reddevils, r/Gunners, r/coys, r/chelseafc	Sports
r/Ladyboys, r/ShemalesParadise, r/ProneSpread, r/amateur_shemales, r/Asian_Ladyboys, r/StandingAssSpread, r/ThaiLadyboy	Pornography
r/GoNets, r/nbacirclejerk, r/lakers, r/sixers, r/bostonceltics	Sports
r/AnimeGirlsAndThongs, r/AnimeGirlsinJeans, r/officelady, r/somuchhentai, r/NikkeNSFW	Pornography
r/GayKnots, r/gayyiff, r/deerbutt, r/gayfurryporn, r/femyiff	Pornography
r/ShingekiNoKyojin, r/attackontitan, r/titanfolk, r/yeagerbomb, r/AttackOnRetards	TV Series
r/LaborPartyofAustralia, r/Australia_, r/PoliticsDownUnder, r/friendlyjordies	Politics
r/amazingtits, r/LiveFreeCams, r/OnlyAmateurPorn, r/Solo_Girls	Pornography
r/IndieDev, r/IndieGaming, r/Unity2D, r/Unity3D	Games
r/GolemSexy, r/BitedSizeSexy, r/ThinkTankSexy	Pornography
r/FeetLoversHeaven, r/feetpics, r/VerifiedFeet	Pornography

Table 2. *Cont.*

Subreddit Group	Topic
r/Maps, r/MapsWithoutNZ, r/ShittyMapPorn	Maps
r/FemboyHentai, r/FemboysAndHentai, r/SabuArt	Pornography
r/gameofthrones, r/HouseOfTheDragon, r/asoiafcirclejerk	TV Series
r/NTR, r/CartoonPorn, r/Cartoon_Porn	Pornography
r/oculus, r/virtualreality, r/OculusQuest	Games
r/PlantedTank, r/aquarium, r/Aquariums	Aquaristics
r/CelebsWithPetiteTits, r/Celebhub, r/CelebrityBelly	Pornography
r/microgrowery, r/cannabiscultivation, r/GrowingMarijuana	Drugs
r/honkaiimpact3, r/Elysia, r/houkai3rd	TV Series
r/BadChoicesGoodStories, r/AmericanFascism2020, r/OliverMarkusMalloy	Politics
r/TransLoves, r/TSfuck, r/TAmateurs	Pornography
r/hentaifemdom, r/gentlefemdom, r/mommydom	Pornography
r/ThunderThots, r/SheGotHands, r/FightsGoneWildpt2	Pornography

The majority of graph-similar subreddit groups are related to various types of pornography. Noticeably, there is a standing-out separate group of subreddits with Indian pornography (subreddits with "Desi" in their names). There are also groups about sports (especially football), politics, and anime. This is very consistent with previous findings on the text content similarity and embedding clustering [14], which also found that the biggest group of subreddits are related to pornography, games, politics, sports, and movies/TV series. The most noticeable niche groups concern aquarium and map-related subreddits, which also have high node embedding similarity in CrosspostNet. These two groups were also noticeable in the previous work [14]. This, again, shows a relation between content similarity and crossposts. Furthermore, the consistency between content similarity and crosspost similarity suggests that crossposts could be used, in studies, instead of (or to cross-validate) content similarity.

5. Selected Results Related to Past Subreddit Studies

In addition to macro-scale characteristics of crosspost-driven subreddit-based graph networks, relations between particular subreddits have been explored. Here, the presented findings provide a background that can be used, in the future, by other researchers working on specific phenomena modeled with subreddits and communities.

5.1. Crossposts Used for Community Conflicts

Reddit is a platform for discussion, but also for various internal conflicts and inter-subreddit quarrels [17]. This has been studied previously [17] and several groups of subreddits were found to take part in "subreddits conflicts", with r/SubredditDrama being the top contributor/host of "drama" related topics. The previous paper analyzed the anti-social behaviors on Reddit, by constructing a conflict network.

These conflicts are realized by regular user activities, such as posting and commenting. In this work, the question as to whether there is any correlation between subreddit conflicts and crossposting activity has been posed. In other words, are crossposts used as a tool in an inter-subreddit conflict? The original paper, explicitly names the highest conflicting subreddits: r/SubredditDrama, r/EnoughTrumpSpam, r/The_Donald, r/politics.

For all these and other subreddits considered in this study, all crossposts that appeared between them have been checked. None of the groups, or subreddit pairs, had a higher proportion of crossposts between them. Next, the number of total crossposts between these

subreddits has been established, and the crosspost score and the number of comments on said crossposts added. Again, nothing showed any anomalies.

What should be noted, is that the original paper analyzed data from 2016. Crossposts were introduced in 2017 (https://www.reddit.com/r/modnews/comments/7a5ubn/crossposting_coming_soon_to_your_subreddit/, accessed on 22 August 2023). So it is impossible to compare the results directly. However, the original paper suggested that the conflict posts and comments do not have a positive sentiment. Therefore, using a natural language model, the sentiment of the crossposts titles has been established. Specifically, DistilBERT [25] fine-tuned on the SST-2 dataset [26] was used (https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english, accessed on 22 August 2023). For a given text, the model returns a value between -1 (negative sentiment) and 1 (positive sentiment). The mean sentiment of all crossposts (used in this study) is -0.14 (std. 0.86), while the sentiment within said group is -0.36 (std. 0.65). Although the mean sentiment is smaller (more negative) within the target group, the standard deviations indicated the enormous variation of sentiment for different subreddit pairs. This may suggest that, as a general rule, crossposts are not used within conflicts or, in general, to deliver negative content on Reddit. Even though individual crossposts may be "emotional". In other words, crossposts are neither used as an offensive nor a defensive tool in the context of subreddit conflicts.

5.2. Relations between Subreddits Previously Analyzed in the Literature

There have been multiple microscale researches on Reddit, studying a particular subreddits, or a group of selected subreddits. We now illustrate how analysis of crosspost-driven graph networks may be useful for researchers interested in specific subject areas. Obviously, the material presented in what follows is just a set of examples that may be generalized into a useful proposal.

Two studies about social norms [27] and social morals [28] considered the subreddit r/AmITheAsshole. Checking the connections of this subreddit, in the crosspost-based graph, it was easy to spot that it is most strongly connected to two other subreddits: r/AmITheAngel (out-crossposts count: 3461, top 1% of all crossposts connections)), r/AmITheDevil (out-crossposts count: 2504, top 1% percentile of all crossposts connections)). Even though these subreddits form the most strongly connected "semi-triangle" (by edge weight) within the graph, the edge between r/AmITheAngel and r/AmITheDevil is missing. This makes r/AmITheAsshole the focal point of the discussion. This can be seen as a strong indicator that research reported in [27] and in [28] could benefit from taking into account not only the r/AmITheAsshole, but also co-analyzing r/AmITheAngel and r/AmITheDevil subreddits.

A study focused on mental health (reported in [29]) analyzed the following subreddits: r/alcoholism, r/anxiety, r/bipolarreddit, r/depression, r/mentalhealth, r/MMFB (Make Me Feel Better), r/socialanxiety, and r/SuicideWatch. Obviously, these are focused directly on various mental health issues. However, applying the proposed approach, it was easy to realize that there is a significant number of crossposts between the studied subreddits and other subreddits, presented in Table 3.

Table 3. Top subreddits connections (by crossposts count percentile) for studies on mental health [29].

To Subreddit	From Subreddit	Crossposts Count Percentile
depression_help	depression	0.969402
mentalhealth	helpme	0.937691
mentalhealth	Advice	0.900146
depression_help	mentalhealth	0.900146
mentalhealth	depression	0.900146
mentalhealth	mentalhealth	0.860143
mentalhealth	disorders	0.860143
mentalhealth	OCD	0.860143

This suggests that the results reported in [29] could be made somewhat more comprehensive if the number of analyzed subreddits was extended by some of these that are closely crosspost-related. Here, note that a clear break point occurs after the first subreddit, next after the second subreddit, and another one after the top five subreddits.

Another study, regarding stress analysis, was reported in [30]. It was based on r/stress, r/anxiety, r/homeless, r/almosthomeless, r/Assistance subreddits. Applying the proposed approach, it was established that these subreddits have a significant number of crosspost connections with subreddits shown in Table 4.

Table 4. Top subreddits connections (by crossposts count percentile) for studies on mental health [29].

To Subreddit	From Subreddit	Crossposts Count Percentile
homeless	antiwork	0.922907
homeless	vagabond	0.900146
almosthomeless	Advice	0.860143
homeless	Assistance	0.860143
bayarea	homeless	0.770256
homeless	LateStageCapitalism	0.770256
homeless	urbancarliving	0.770256
almosthomeless	povertyfinance	0.770256
homeless	Portland	0.770256
almosthomeless	homeless	0.770256
homeless	vancouver	0.770256

The results presented in Table 4 show, again that the dataset used in [30] could have been expanded at least by the first two subreddits (after which a clear break can be spotted).

Yet a different study was devoted to parenting communication [31]. It used r/Mommit, r/Daddit, and r/Parenting. Applying the proposed approach, it could be suggested to also take a look at subreddits in Table 5 as they are closely crosspost-related to the ones used in the original study. Here, the first two or the first five subreddits could be incorporated.

Table 5. Top subreddits connections (by crossposts count percentile) for studies on mental health [29].

To Subreddit	From Subreddit	Crossposts Count Percentile
beyondthebump	Mommit	0.947961
Mommit	beyondthebump	0.922907
Mommit	WhitePeopleTwitter	0.900146
Mommit	aww	0.900146
toddlers	Parenting	0.900146
Mommit	NonPoliticalTwitter	0.860143
Mommit	meirl	0.860143
Mommit	oddlysatisfying	0.860143
Mommit	nextfuckinglevel	0.860143
Mommit	FunnyAnimals	0.860143
parentsofmultiples	Mommit	0.860143
Mommit	comics	0.860143
Mommit	ukraine	0.860143
AttachmentParenting	Mommit	0.860143
workingmoms	Mommit	0.860143

Finally, a political science research was reported in [32]. It was focused on r/The_Donald, r/Conservative, r/politics, r/SandersForPresident subreddits. When crossposts are considered, this set of subreddits could have been expanded by subreddits presented in Table 6.

Table 6. Top subreddits connections (by crossposts count percentile) for studies on mental health [29].

To Subreddit	From Subreddit	Crossposts Count Percentile
Political_Revolution	politics	0.999927
uspolitics	politics	0.999907
conservatives	Conservative	0.999875
The_Mueller	politics	0.999845
FuckGregAbbott	politics	0.999803
Republican_misdeeds	politics	0.999799
politics	sanepolitics	0.999781
Kossacks_for_Sanders	politics	0.999638
TexasPolitics	politics	0.999449
politics	VoteDEM	0.999211
NewPatriotism	politics	0.999143
Fuck45	politics	0.998977
usa	politics	0.998865
ArizonaLeft	politics	0.998689
SandersForPresident	NewDealAmerica	0.998668
RepublicanValues	politics	0.99862
texas	politics	0.998118

However, note that here a rather large group of very tightly connected subreddits has been spotted. It is not our role to decide which ones should (or should not) be included in an extended study. The point is that, from the crosspost perspective, the set of closely related subreddits is larger than those used in the original study.

In summary, we stress that the aim of this section was not to criticize the results presented in mentioned works. Instead, it was to show that when a crosspost-driven subreddit-focused graph network view on the information structure of Reddit is applied, additional subreddits (closely related to these that were originally selected) may be suggested. In other words, the following approach may be suggested.

1. For a given topic of interest, select one or more subreddits.
2. In the case when more than one subreddit is selected, use the proposed approach to verify if these subreddits are connected (and how strongly).
3. Establish which other subreddit(s) are crosspost-connected to the selected ones (and how strongly).
4. On the basis of this analysis, adapt (or do not change) the subreddit selection for the research that is to be undertaken.

Obviously, methods based on links between users (refs. [1–3]) could be also used here. Moreover, both approaches could be combined in this stage of preparatory work. In this way, the "best" set of subreddits could be selected to study the specific topic of interest.

6. Conclusions

In this study, relations between Reddit's subfora were modeled using a graph network-forming algorithm based on crossposts. While the overall characteristics of the network showed results similar to those obtained in previous studies, there have been also new discoveries. When modeled using crossposts, the space of subreddits is also subject to the power distribution of attention and small-worlds phenomena. There are a couple of "dominating" subreddits such as r/HolUp, r/Unexpected, r/interestingasfuck, or r/damnthatsinteresting—all having the most crossposts and all presenting a very generic content. This content can be seen as material that is supposed to amaze the viewer. Moreover, there are many small communities (graph components and communities discovered using the Louvain algorithm) of subreddits with meaningful crosspost relations, but they are of a minuscule scale compared to the top ones. Additionally, the network structure algorithms were used to extract particular relations between subreddits, to suggest and enhance the focus groups of previous studies on social, medical, and political aspects. This showed a useful and practical application of the crossposts network. which can be used

in the initial stages of research related to any topic that is discussed within subreddits. Overall, the findings are consistent with previous analyses of link-based networks, as well as the textual embedding similarity. It can be thus stated that they cross-validate each other and bring more certainty to each one of them, considered separately. In the future, it is worth exploring the CrosspostNet on a larger scale and automating the extraction of similar subreddits using the crosspost connections. Moreover, the evolution of graph networks, related to the strength of crosspost links can be explored. Finally, the time-based evolution of graph networks should be explored to establish if the results presented in this work are the same as in the previous year(s) and/or in the following year(s).

Author Contributions: Conceptualization, J.S. and M.P.; methodology, J.S.; software, J.S.; validation, M.P. and M.G.; formal analysis, J.S.; investigation, J.S.; resources, M.G.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S., M.P., M.G. and Y.W.; visualization, J.S.; supervision, M.P., M.G. and Y.W.; project administration, M.P.; funding acquisition, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baowaly, M.K.; Kibirige, G.W.; Singh, B.C. Co-comment network: A novel approach for construction of social networks within reddit. *Comput. Sist.* **2022**, *26*, 311–323.
2. Steinbaur, T. Information and social analysis of Reddit. *Proc. Troysteinbauer CS. UCSB. EDU*, 2012; pp. 1–12. Available online: https://api.semanticscholar.org/CorpusID:17491485 (accessed on 22 August 2023).
3. Buntain, C.; Golbeck, J. Identifying social roles in reddit using network structure. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014; pp. 615–620.
4. Stewart, L.G.; Spiro, E.S. Nobody puts redditor in a binary: Digital demography, collective identities, and gender in a subreddit network. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 8. [CrossRef]
5. Kwon, K.H.; Yu, W.; Kilar, S.; Shao, C.; Broussard, K.; Lutes, T. Knowledge sharing network in a community of illicit practice: A cybermarket subreddit case. In Proceedings of the 53rd Hawaii International Conference on System Sciences, Maui, HI, USA, 7–10 January 2020.
6. Hurtado, S.; Ray, P.; Marculescu, R. Bot detection in reddit political discussion. In Proceedings of the Fourth International Workshop on Social Sensing, Montreal, QC, Canada, 15 April 2019; pp. 30–35.
7. Chipidza, W. The effect of toxicity on COVID-19 news network formation in political subcommunities on Reddit: An affiliation network approach. *Int. J. Inf. Manag.* **2021**, *61*, 102397. [CrossRef] [PubMed]
8. Yoo, M.; Lee, S.; Ha, T. Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. *Inf. Process. Manag.* **2019**, *56*, 1565–1575. [CrossRef]
9. Del Valle, M.E.; Gruzd, A.; Haythornthwaite, C.; Kumar, P.; Gilbert, S.; Paulin, D. Learning in the wild: Predicting the formation of ties in 'Ask'subreddit communities using ERG models. In Proceedings of the 11th International Conference on Networked Learning, Zagreb, Croatia, 14–16 May 2018; pp. 157–164.
10. Krohn, R.; Weninger, T. Subreddit Links Drive Community Creation and User Engagement on Reddit. In Proceedings of the International AAAI Conference on Web and Social Media, Limassol, Cyprus, 5–8 June 2022; Volume 16, pp. 536–547.
11. Medvedev, A.N.; Lambiotte, R.; Delvenne, J.C. The anatomy of Reddit: An overview of academic research. In *Dynamics on and of Complex Networks III: Machine Learning and Statistical Physics Approaches 10*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 183–204.
12. Onnela, J.P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D.; Kaski, K.; Kertész, J.; Barabási, A.L. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7332–7336. [CrossRef] [PubMed]
13. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47. [CrossRef]
14. Sawicki, J. Text embeddings and clustering for characterizing online communities on Reddit. In Proceedings of the 18th Conference on Computer Science and Intelligence Systems (FedCSIS), Warsaw, Poland, 17–20 September 2023.
15. Dijkstra, E.W. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy*; ACM Digital Library: New York, NY, USA, 2022; pp. 287–290.
16. Takes, F.W.; Kosters, W.A. Computing the eccentricity distribution of large graphs. *Algorithms* **2013**, *6*, 100–118. [CrossRef]
17. Datta, S.; Adar, E. Extracting inter-community conflicts in reddit. In Proceedings of the International AAAI Conference on Web and Social Media, Munich, Germany, 11–14 June 2019; Volume 13, pp. 146–157.

18. Van Pham, H.; Tien, D.N. Hybrid louvain-clustering model using knowledge graph for improvement of clustering user's behavior on social networks. In Proceedings of the Intelligent Systems and Networks: Selected Articles from ICISN 2021, Hanoi, Vietnam, 19 March 2021; Springer: Singapore, 2021; pp. 126–133.
19. Clauset, A.; Newman, M.E.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*, 066111. [CrossRef] [PubMed]
20. Edler, D.; Bohlin, L.; Rosvall, M. Mapping higher-order network flows in memory and multilayer networks with infomap. *Algorithms* **2017**, *10*, 112. [CrossRef]
21. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
22. Salehi Rizi, F.; Granitzer, M. Properties of vector embeddings in social networks. *Algorithms* **2017**, *10*, 109. [CrossRef]
23. Grover, A.; Leskovec, J. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–16 August 2016; pp. 855–864.
24. You, X.; Ma, Y.; Liu, Z.; Liu, J.; Zhang, M. Representation method of cooperative social network features based on Node2Vec model. *Comput. Commun.* **2021**, *173*, 21–26. [CrossRef]
25. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
26. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.; Potts, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
27. De Candia, S.; De Francisci Morales, G.; Monti, C.; Bonchi, F. Social norms on reddit: A demographic analysis. In Proceedings of the 14th ACM Web Science Conference 2022, Barcelona, Spain, 26–29 June 2022; pp. 139–147.
28. Botzer, N.; Gu, S.; Weninger, T. Analysis of moral judgment on reddit. *IEEE Trans. Comput. Soc. Syst.* **2022**, *10*, 947–957. [CrossRef]
29. De Choudhury, M.; De, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 71–80.
30. Turcan, E.; McKeown, K. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv* **2019**, arXiv:1911.00133.
31. Sepahpour-Fard, M.; Quayle, M. How do mothers and fathers talk about parenting to different audiences? Stereotypes and audience effects: An analysis of r/Daddit, r/Mommit, and r/Parenting using topic modelling. In Proceedings of the ACM Web Conference 2022, Barcelona, Spain, 26–29 June 2022; pp. 2696–2706.
32. Soliman, A.; Hafer, J.; Lemmerich, F. A characterization of political communities on reddit. In Proceedings of the 30th ACM Conference on Hypertext and Social Media, Hof, Germany, 17–20 September 2019; pp. 259–263.