UNIVERSITY OF TARTU
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Computer Science

Software Engineering

Oliver Soop

Local Information Diffusion Patterns in Social and Traditional Media: The Estonian Case Study

Master's thesis (30 ECTS)

Supervisor: Peep Küngas, Ph.D.

TARTU 2014

Local Information Diffusion Patterns in Social and Traditional Media: The Estonian Case Study

Abstract

Information has become more highly valued among companies and individuals than ever before. With this, the interest in how information diffuses among the entities in various structured networks has increased. A number of studies have been published on the diffusion process in real-life networks, such as web service network, citation networks, blog networks etc. Majority of researches have focused on one type of network - such as Facebook posts, Twitter tweets, Blogspot blog entries etc. A disadvantage of analysing a network containing entities from a single source is that it does not consider the outside influence on the diffusion. Recently, some papers have started to incorporate different networks in their study and as such have been able to analyse the effect of outside influence on the diffusion process.

This thesis aims to shed further light into the topic of information diffusion in a real world network containing entities from different sources, this is achieved by detection of relevant local topological and temporal information diffusion patterns. For topological pattern analysis, frequent subgraph mining techniques are used. Temporal patterns are extracted using time series clustering. The dataset used in this thesis is collected from the Estonian setting of mainstream online news media with comments and articles and from social media channels Twitter and Facebook. From this dataset the relations between the entities were extracted and a network for analysis of diffusion patterns was constructed. Temporal patterns reveal the high pace of information diffusion while topological patterns expose the important role of news media articles and Facebook posts in the information diffusion processes. The results of the thesis are applicable in cyber defence, online marketing and campaign management plus information impact estimation, just to mention a few application areas.

Keywords:

information diffusion patterns, information diffusion, social media, traditional media

Kohalikud informatsiooni levimise mustrid sotsiaal- ja tavameedias: Eesti kontekst

Lühikokkuvõte

Paljud ettevõtted ja inimesed hindavad kõrgelt informatsiooni väärtust ja seda on eelkõige hakatud hindama viimase kümnekonna aasta jooksul. Tänu sellele on tekkinud ka huvi, kuidas info levib erinevates struktureeritud võrgustikes. Avaldatud on mitmeid teadustöid, mis uurivad informatsiooni levimist ühes reaalse elu võrgustikus nagu näiteks Facebooki postitused, Twitteri tweetid, Blogspoti blogikanded jne. Suuresti on need uurimused keskendunud ühele võrgustikule, mis ei hõlma kogu võrgu dünaamikat ja samuti välist mõju info levimisele. Samas on lähiminevikus avaldatud ka

teadustöid, mis hõlmavad mitut erinevat võrgustiku ja analüüsivad välist mõju informatsiooni levimisele.

Käesoleva töö eesmärk on lähemalt uurida informatsiooni levimise mustreid võrgustikus, mis hõlmab erinevaid reaalelu võrgustike, kasutades selleks topoloogilisi ja aja mustreid. Topoloogiliste mustrite analüüsimiseks on kasutatud võrgustikus sagedalt levivate alamgraafide leidmise algoritme, aja mustreid uuritakse ajaseeriate klasterdamise teel. Töös kasutatud andmestik on kogutud Eesti uudismeediast - artiklid ja nende kommentaarid ning sotsiaalmeedia kanalitest, Twitterist ja Facebook-ist. Selle andmestiku põhjal loodi seosed eritüüpi andmeobjektide vahel, mille põhjal loodi võrgustik, mida kasutada edasiseks uurimiseks. Aja mustrid viitavad väga kiirele info levimisele antud võrgustikus, topoloogilised mustrid näitavad uudismeedia artiklite ja Facebook-i postituste suurt mõju info levimises. Töö tulemusi on võimalik rakendada küberkaitses, online turunduses ja kampaania haldamises, samuti ka mõjuvõimu hindamisel - kindlasti leiaks tulemused rakendust ka teistes valdkondades.

Võtmesõnad:

informatsiooni levimise mustrid, informatsiooni levimine, sotsiaalmeedia, uudismeedia

Contents

1 Introduction

Information diffusion as a concept has not been precisely defined, but one description of this process is the act of an entity being exposed to some information and subsequently, the entity exposes it to other related entities using different communication channels. The rise of the amount of data created constantly and the emergence of large social networks has created even more interest in information diffusion. Gantz and Reinsel study [13] reports an estimate that the amount of data created yearly will grow 300 times from 130 exabytes in 2005 to 40000 exabytes by 2020. This means that the quality of data and more precisely, the quality of information will become increasingly important. One of the ways to quantify the quality of information is also by exploring the characteristics of information cascades. The genesis of social networks like Twitter, Facebook, Youtube, Flickr etc. has also created a new channel for receiving daily news and information. Compared to traditional online news media channels these social networks provide their users with opportunity to have a more concentrated information feed. As a result, the process of information diffusion has also become more ambivalent and is mostly guided by the related entities in one's network. These networks are also fascinating from the aspect of their size; according Fowler [12] in October 2012 Facebook had an estimate of about 1 billion users. Twitter was reported to have passed the 500 million user count in July 2012 as stated in Semiocast publication [34]. With the rise of the amount of data, quite interesting phenomena have also appeared: faster expiration of information, rise of the value of novel and significant information, measure of influence of members of social network. Additionally the social networks have brought up the aspect of rapid diffusion of information among the people – viral effect. This is mainly the subject of marketing campaigns generated by different companies trying to get the information about their product out to the largest number of people possible.

Several studies have been performed on information diffusion in real life networks, such as social networks by Gomez-Rodriguez et al. [15] and Yang and Leskovec [44], also about the blog network structure by Cui et al. [8] and Yang and Counts [46]. The studies have researched the diffusion process from a variety of perspectives – some of them have tried to create descriptive models of the diffusion processes and some have researched information diffusion by trying to come up with models that help to predict future information diffusion. One example of descriptive modelling is the NETINF algorithm by Gomez-Rodriguez et al. [15]. From practical viewpoint Shahaf et al. [35] created multi-resolution metro map of a news topic diffusing across multiple news media channels over a period of time - one of their relevant results was discovering diffusion cascades based on the contents of the news items. There are also more specific studies that concentrate on various features of information diffusion, for example Stewart et al. [37] were interested in diffusion paths for more efficient online advertising. Budak et al. [6] were searching solutions for how to counter the diffusion process of rumours and misinformation.

Most of these studies have been conducted with the assumption of a closed network – so that they present all the characteristics for describing the dynamics and diffusion patterns of the networks that are under observation. But this does not cover the word-of-mouth implications that are subject to occur, also there are other numerous aspects that have not been incorporated and may also characterise the diffusion process. These studies of closed networks mostly incorporate only one specific

domain, may that be the citation network, social network, blog network or something else. Therefore they do not reflect the outside influence that may lead the diffusion process in some circumstances like, for example, during an emergency event and the information propagation in news media.

Recently some studies have been done that also research the external effect on information diffusion - one example of this is the study by Myers et al. [30] on the influence of mainstream media on the information diffusion processes in social media. These studies have been done in global scale and have resulted in general estimates of the degree of influence.

In this thesis, information diffusion process will be researched in a network that embodies social media and traditional media domains and is not limited to only one social media network. The dataset used in this thesis is based on the Estonian traditional online media and social media. First, a single network will be created that is based on online media news articles and comments, Twitter users' tweets and Facebook posts, commentaries and likes. The relations between different entities of the network will be constructed on the basis of different features of the networks used. For Twitter one of the patterns for describing relations between two entities or tweets is the use of a retweet mechanism. Another approach that will help to determine relations between entities of different domains is the exploration of Uniform Resource Locators (URL) that refer one entity from the other. URL references are explored in every domain and this is the basis for the creation of an interrelated network of the diverse networks.

Kwak et al. [22] have raised the question whether Twitter actually belongs to the social media domain, by showing that the usual characteristics that best describe social media tend not to conform to Twitter network. As well Sakaki et al. [33] reveal the nature of Twitter as a real-time event information diffusion network - displaying evidence that news about current events (e.g. earthquakes) can be detected very early in Twitter network. As the division of the types of nodes in the network is defined by the descriptive words of traditional online news media and social media and the content in Twitter is collective creation of different people rather than the necessity of their job description - here Twitter is still considered to be part of social media.

After the data cleaning and creation of a time dimensional network, the graph will be studied and analysed using different graph metrics like centrality measures, vertex degree evaluation. Graph research tools and methods will be used for discovering temporal and topological information diffusion patterns. Information diffusion patterns will be explored to discover how information flows in one domain, how it crosses borders between different channels and also between domains, creating information diffusion cascades. Information cascades and the patterns these cascades depict will be examined from two perspectives: temporal and topological viewpoint. From temporal perspective, the patterns describe how much time it takes for information to flow and when certain events take place in social and traditional media. First time series are extracted and clustered using k-means algorithm - most common temporal patterns are analysed. Temporal perspective can reveal if there are bottlenecks during any type of information propagation and what is the expected time for the ongoing propagation to decay. Viral diffusion process is another phenomenon that can become vivid when studying temporal patterns of the diffusion cascades. With frequent subgraph mining

techniques relevant information diffusion patterns available in the network will be extracted and these topological patterns will be analysed. Particularly interesting is the information flow between social media and traditional media domain - also if there is any evidence of some entities that initiate most of the cascades. The significant patterns will be evaluated and analysed based on the viewpoints discovered in other relevant studies with the goal of identifying their importance in the network - furthermore, the role of different types of vertices in the diffusion process will be analysed. Lastly, conclusion will be drawn upon the analysis.

The thesis is structured as follows: the second chapter contains a current reflection of the studies that have been done in this field. The third chapter presents the background and methods that were used to discover temporal and topological diffusion patterns and to characterise the diffusion process. In the fourth chapter, the initial dataset and its properties are introduced as well the data extraction process with the graph construction is described. The fourth chapter also gives an overview of the network by analysis of different graph metrics. In the fifth chapter, the network is analysed from the temporal perspective along with the analysis of temporal patterns. Second part of the fifth chapter analyses most common topological patterns. In this chapter, the analysis tries to answer two research questions: "What kind of temporal information diffusion patterns emerged?" and "What are the different common topological patterns available in the network?". The sixth and also the last chapter concludes the thesis and describes possible future work.

2 Related Work

Information diffusion has been studied in various contexts. Specific studies include for instance diffusion in web services networks [Mokarizadeh et al. – [29]], social networks [Gomez-Rodriguez et al.[15], Yang and Leskovec [44], Taxidou and Fischer [38]], citation networks [Shi et al. [36]], blog networks [Cui et al. [8], Yang and Counts[46], Kwon et al. [24]], multi-domain networks [Myers et al. [30], Kim et al. [21], Kwon et al. [23]] etc.

Different algorithms and approaches have been explored in measuring and researching information diffusion. In general, research around this problem has been conducted mainly by defining the information diffusion problem in terms of a network analysis problem and then, specific network characteristics are measured to reveal the patterns of diffusion. This approach has been primarily exploited in social network analysis (SNA). One example of SNA is the NETINF algorithm by Gomez-Rodriguez et al. [15] that was developed to infer networks of information diffusion and influence. This algorithm is a novel method that is based on studying the cascades of nodes getting infected instead of the source of infection, such that root cause analysis can be performed on the propagation traces from which the network can be deduced. This algorithm has merits in settings where the source of infection is not explicit such as in networks that span across multiple domains, for instance, the blogosphere and mass media, which was the case in the paper by Gomez-Rodriguez et al. [15]. This study concentrated on the analysis of diffusion patterns in blogosphere and mass media and revealed that mass media forms a core-periphery structure that influences majority of the blogosphere.

Yang and Leskovec [44] describe in their study a model of diffusion in the absence of explicit knowledge about the source of the infections in the network. Their solution is a Linear Influence Model that is based on the assumption that the number of nodes that get infected depends on which nodes got infected in the past. In this model an influence function is associated with every node. One of the findings reported in the study is that in Twitter people who have the most followers do not tend to be the most influential when it comes to information propagation.

Compared to the approaches of Gomez-Rodriguez et al. [15] and Yang and Leskovec [44] in this thesis, we do not assume any explicit knowledge about the source of the "infection" and as a matter of fact we only concentrate on the "infections". Similarly to the paper of Yang and Leskovec [44], this thesis concentrates on mainstream media, blogs and social media, although in the social media perspective we consider additionally Facebook and comments in traditional media.

Although this thesis is based on a finite set of data, one problem in information diffusion lies in the amount of data that is created during a short period of time. As the amount of information becomes increasingly large – the freshness of the data quickly degrades. This issue is tackled by Taxidou and Fischer [38] in the context of analysing information diffusion in real time. The proposed approach is based on the information cascades composed of the stream of messages and evaluations about the messages. Given that the study is based only on the underlying social network and relations in this network - the approach does not study how the initial "infection" in this network takes

place. However, the source of the initial infection may provide valuable information about how the information propagates through the network.

The aspect of influence and information diffusion not being constrained in the borders of a single social network or media source has been studied by Myers et al. [30]. One of the key findings in the paper is that considerable number of mentioned URLs (29% of Twitter URLs as reported in the study) are created due to external influence. In this research paper, the authors apply the notion of two states of the diffusion process - exposure and infection. External influence is modelled in terms of these two states. External exposure volume is based on event profile that is inferred by authors from the set of node infections in the network.

Similarly to the study of Myers et al. [30], diffusion process across multiple networks has been studied by Kim et al. [21] by concentrating on the event diffusion in and between social networks like Facebook, Twitter and Flickr; news media (e.g. NY Times, BBC) and blogosphere (BlogSpot, Wordpress etc.). The dataset used in this thesis is created by traversing URLs between documents. The network structure contains in addition to links extracted from document URLs, as done in majority of related studies, also links based on the similarity of different document contents. The network considered by authors is a bipartite graph containing documents and users contributing to those documents. One of the findings reported in this study is that interaction between social network entities and news media entities is unidirectional while the interaction between news media and blogosphere is bidirectional. Another finding is that 1% of most productive users produce over 40% of the content related to the corresponding event.

Unlike this thesis Myers et al. [30] does not incorporate different networks into a holistic view in their study where they use a probabilistic generative model to evaluate external influence on information diffusion. With respect to the perspective of the dataset, the paper by Kim et al. [21] is based on different networks and only concentrates on important political, economic, disaster and sport events.

There are numerous articles that have studied information propagation in networks and significant number of them concentrate on a single static network. One of the most studied networks is the Twitter network, which is a micro-blogging system. For example, Kwon et al. [23] study information diffusion with the emphasis on comparing the content that crosses over from other social media sites to that when the content is internal to the network. Experiments conducted by Matsubara et al. [27] were also partly based on the Twitter network. The main research goal of the authors was to create a novel method to forecast the event diffusion in time. For this, the authors created the SPIKEM analytical model to model the rise and fall patterns of influence propagation. Myers et al. [30] also based their work on the Twitter network.

There are other micro-blogging systems as well that have been studied with respect to information diffusion. For example, Cui et al. [8] concentrate on Sina Weibo, a Twitter-like micro-blogging system to study the mass media spread mechanism in social media during emergency events. The authors compare the information diffusion process in the social network to the word-of-mouth dif-

fusion process.

Besides micro-blogging networks, there are many research papers where studies are centred around blogosphere. For instance, Yang and Counts [46] study information diffusion structure in both micro-blogs and weblogs. As a result, the authors outline the key differences between Twitter and weblogs from the perspective of contribution, navigation and network characteristics. One divergence found is that the structures of the networks differ in great amount – the Twitter network is more decentralized and connected locally, while the weblog network is more coherent globally. In relation to this thesis, the authors point out in their research that most of the URLs posted in Twitter are outbound – this shows that the network can be well enhanced with the addition of other domain networks. Kwon et al. [24] report a study fully concentrated on a blog network. More specifically, the authors study diffusion in the Korean blog scene. The paper concludes that most of diffusion does not take place between the nodes that have an already established relation. In the context of online advertising Stewart et al. [37] have designed a method for mining information diffusion paths in blog networks. The solution is an algorithm that is based on frequent pattern mining and other algorithms like FP-growth and the FS-miner that were regarded most suitable for this kind of frequent itemset and sequence mining. Lim et al. [26] develop a method for the construction of a blog network with the assignment of diffusion probabilities to each node. They try out their solution on the Korean blog service of blog.naver.com.

After the creation of Facebook, one of the biggest social networks currently, numerous research papers have been published studying the information diffusion behaviour in this context. Bakshy et al. [4] study shows that exposure to information makes entities to more likely spread information and do this sooner. They also find that weak links between entities contribute to the dissemination of novel information. They also bring out the ambiguity of the diffusion of information among peers – arguing that it may as well be the result of a peer effect or common interests. Xu and Liu [43] study diffusion process in Facebook with the intention of preventing propagation of rumours. They improve the Susceptible-Infected-Susceptible model which is used to model dynamics of diffusion among social groups by covering the shortcoming of the model when it comes to the decaying of the topic. Further they improve this model by employing the factor of the stem of new subtopics in the propagation process. Budak et al. [6] evaluate their research of limiting the spread of misinformation in social networks on a dataset gathered from Facebook. They develop a method for counter measuring the diffusion process of information that is not valid by identifying subset of influential entities to target for starting a opposing campaign.

An overview of different research questions and current methods and models for studying information diffusion in online social networks is given by Guille et al. [17]. They concentrate on the current state in identifying relevant topics, descriptive and predictive modelling of information diffusion and also cover the aspect of influence and influential spreaders. Guille et al.[16] have also publicised a platform for mining and analysing social networks. The tool, called SONDY (Social Network Dynamics) provides functionality for data manipulation and cleaning, it also includes methods for analysis of influence and visualization of the results.

3 Dataset

The approach taken to extract the local information diffusion patterns in social and traditional media includes the following steps: data extraction, network construction, data cleaning, network analysis. An overview of the workflow of this thesis is given in Figure 1. In more detail - at first the initial data available will be studied, based on this the diffusion relations will be explored and chosen. Diffusion relations fall into two categories: explicit relations - ones that were available in the dataset, for example Twitter and retweeting, Facebook post and commenting, implicit relations - URL references to other entities in the different datasets. Based on the amount of data in initial datasets a temporal data extraction plan will be created and implemented accordingly. As a result of this phase, intermediary datasets will be created upon which the graph could be constructed.

Next step in this process will be graph construction - for this all of the intermediary datasets will be combined and URL based relations discovered. Combining the datasets means that the entities in datasets will be reindexed so that everything is based on one index namespace. After constructing the graph it will be explored and as a result the shortcomings of the data extraction and also graph construction processes will be visible to be analysed and dealt with. Therefore, dealing with any of the biases found data cleaning will follow next.

When the final graph is available it will be studied using different metrics - most of the metrics are chosen based on different studies that have found them to be relevant. These studies have analogous research topics - information diffusion, social media, traditional media etc. A few of the metrics, like degree distribution, centrality measures give a good overview of the topology of the graph and thus of the information diffusion in social and traditional media. Comparing these values to the results of other studies gives useful insight on what kind of similarities exist between the constructed network and networks under study in other researches.

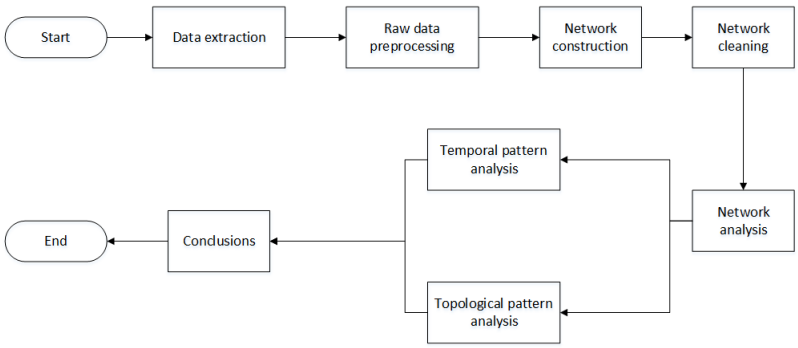


Figure 1: Workflow taken to reach the final goals

3.1 Initial Dataset

The network under study was created by extracting data from different datasets. These datasets were provided by Estonian company Register OÜ, this company provides information services about Estonian companies in a structured manner. These datasets include whole web pages of traditional

online news media; the comments that were given to the news items; Facebook posts, likes, comments; Twitter tweets. The full web page news media dataset also includes extracts from blog RSS feeds. Therefore the created network will incorporate data from multiple different domains – social media, traditional media and article comments. To bring out a few examples from the mainstream online news media – articles from postimees.ee, aripaev.ee, epl.ee are included. All of this data is geographically related to Estonia and Estonian companies – a few discrepancies from the expected geographical location may also be included. Facebook posts, comments and likes were collected from the public Facebook pages of Estonian companies.

News articles that were published by Estonian online media companies were stored as full web pages and therefore would contain date of publishing, author and the content. This also means that there are many sources for establishing relations to other entities. As opposed to articles their comments were stored in a more structured way and in database from where the required information could be queried – for example content of the comment and creation date. Facebook also provides a lot of information, although, usually the length of the content is not that long – may it be a post or a comment to that post. In initial dataset, we have access to posts of the Facebook pages, the commentaries connected with these posts and also the likes relevant to these posts. Twitter dataset is as well stored in a structured manner captured using Twitter API, containing 140 character length post content, author, timestamp and also connections between posts in different feeds.

3.2 Data Extraction

For the creation of a graph in edge list format, the data provided was processed using different techniques - mostly techniques for extracting specific text from content were applied. In the final graph, there are entities (nodes) from two different domains - online media and social media. In more detail the nodes can be of the following type: Facebook post, Facebook comment, Facebook like, Twitter tweet, article and article comment - these node types and their relations are represented in Figure 2. Articles also include the entities of the blogosphere.

- Facebook post - also called Facebook status, is a feature that allows users or companies to post and share textual or other type of content, usually not that long [39].
- Facebook comment - is a textual input so that people can comment on content on a page [10].
- Facebook like - is a representation of clicking a button and sharing pieces of content on the web or Facebook pages with friends [11].
- Twitter tweet - is a message posted on Twitter that is up to 140 characters long [41].
- Article - here we consider online media articles that discusses current events or some specific topic.
- Article comment - textual input or opinion given by people to articles content.

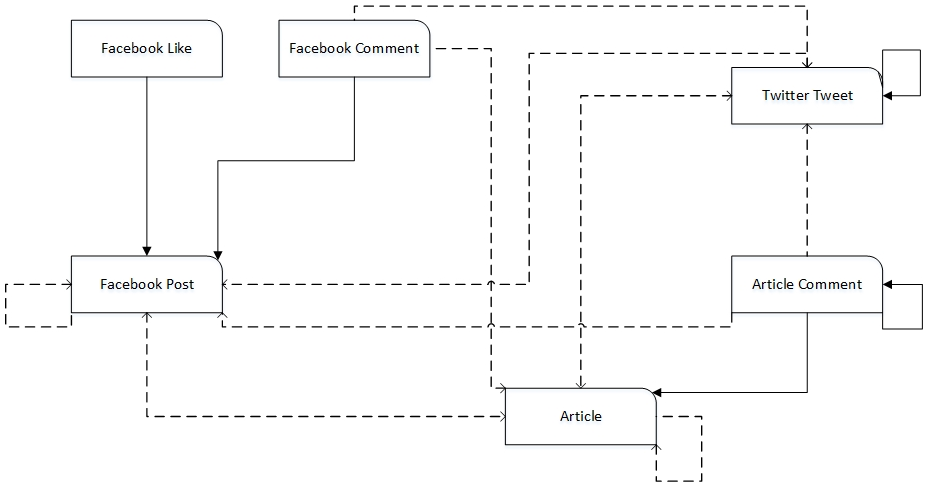


Figure 2: Domain model of different types of nodes. Solid lines represent explicit relations between entities. Dashed lines represent relations found using URLs.

Various methods were applied to construct the edges between the entities of the graph - in some situations the relation between two entities is explicit and easily retrievable from the initial dataset. This is mainly subject to entities that are from the same domain for example the relation between Facebook entities is explicitly provided by the Facebook API. For connecting nodes that are from different domains Uniform Resource Locators (URLs) were used. URLs also explicitly refer to an entity but backtracking from the referred resource is not possible. Therefore the corresponding matching resource had to be found from the dataset to produce a coherent graph with entities from different channels. Data extraction can be largely divided into three phases:

1. Extraction of identifiers and URLs.
2. Validation of URLs.
3. Construction of a graph in edge list format.

The phases themselves were different for all the channels due to the diverse initial datasets and also due to how the data was presented. This entailed various algorithmic approaches to be taken for the extraction and resulted in largely varying time efficacy, therefore requiring algorithm validation beforehand.

3.2.1 Extraction of identifers and URLs

During the first phase the primary goal was to extract URLs that would be later used to create relations between nodes that were of different channels. The news media dataset was the first to be scoured for URLs. Initial dataset for news media was in the format of MySQL dump files and included a table that withhold identifier for the entity and the full extracts from the web pages.

These web pages were articles of news media sites or blog posts. There were approximately 50 of this kind of input files varying in size from 10 GB to 150 GB. To process these files and search for URLs scripts written in Python [1] were made that are available in Appendix 1. To extract URLs from web page excerpts the algorithm included regular expressions that matched for different values in HyperText Markup Language (HTML) files. Listing 1 shows a common example of representing a link that has an URL referencing some resource.

Listing 1: Example URL representation in HTML.

```
<a href="http://www.google.com/analytics">Google Analytics</a>
```

All the URLs in HTML are not always links so furthermore two other strategies were used. Firstly, matching for the World Wide Web acronym www. in the extracts and secondly matching for a subset of country code top-level domains like .ee, .com, .uk, .org etc. The full regular expression for extracting the URLs from HTML document extracts is available in Listing 2.

Listing 2: Regular expression for the extraction of URLs.

```
(?i)((?:<a)(?!/>)(?:[^>]+)(?:href=\")(?P<first>[a-zA-Z0-9/.]+)(?:\"))|
(?P<second>www.[^\s<;\"]+)|
(?P<third>([a-zA-Z0-9.]+)
(\.ee|\.com|\.uk|\.de|\.org|\.fi|\.net|\.lv|\.fr|\.lt|\.us|\.dk|\.biz|\.tk)
([a-zA-Z0-9/.\?]+))
```

Although this will also produce false positives such as “evidently.commander in chief” which is not an URL but actually typing error. In later phases, these false positives will become irrelevant as a matching referenced resource will be sought for and if it is not an URL, none will be found. This will also result in that the URL will be discarded from the final edge list.

The input Facebook data was gathered in a different manner using the Facebook API, and for this reason it was also available in more structured form. Hence the extraction of significant data was easier - relations between Facebook posts, comments and likes were explicitly available in this dataset. The textual content of Facebook posts or comments nevertheless had to be processed to find for any URLs present in them. For this the same approach was taken as described before - regular expressions were used to extract URLs. The script used for the extraction of relations between different Facebook entities and mining of URLs in posts and comments is available in Appendix 2.

Similarly to Facebook data, Twitter input data had also been crawled using Twitter API. Twitter input data contains text message called tweets, published by users. Accordingly, the data was available in a structured manner, although the aspect of retweeting [40] representing a relation between two entities was not explicitly attainable from the input data resources. Only the target users identifiers of the retweet was available and the text of the retweet - therefore, the text of the

retweet were matched to every tweet of the specified target users to find the tweet identifier and confirm relation. For Twitter, there was no need to mine for URLs as these were already made available by the public Twitter API and were also present in the Twitter input data for this thesis. The script written in Python for the extraction of URLs and relations between Twitter tweets is available in Appendix 3.

Article comments were another entity that needed to be processed for URLs that might be contained in the texts. Relations among different comments can exist as well, and these are usually reply-relations - one comment is written in reply to another. These relations were explicitly available in the dataset, only URLs were mined according to the same principles as before - matching regular expression in the comment text. Script for this is available in Appendix 4.

3.2.2 Validation of URLs

Validating URLs was one of the most exhausting process of the three phases used to extract the graph. As Twitter embodies a rule of no tweet being longer than 140 characters this has resulted in the use of the concept of URL shortening. URL shortening service providers give the chance to create a redirect URL that is shorter in length and will redirect you to the initial web page. Short URLs are not only used in tweets but can also be found in other entities. For this reason every URL found in previous phase was subsequently again processed to follow the redirection if it was present. URL validation was only done for URLs that were found in Twitter tweets and Facebook posts and comments. Accordingly short URLs were replaced by the URL that was the result of the last redirection. Incorrect resource locators were not removed from this dataset these include URLs that are not valid or have expired meaning that the referenced resource is not available anymore.

For this phase, another Python script was created that is available in Appendix 5. The script is rather simple by making HTTP requests to the input URL and returning the URL of the response message as being the valid one if the response code is 200 OK [9]. If the response code of the request is not 200 then an assumption is made that the initial resource locator is correct. Including URLs that do give the desired response code are still preserved as it may be that these are expired, but were valid before. This means that in the next phase during the edge list creation a corresponding resource may still be found as the datasets were crawled concurrently.

3.3 Network construction

Edge list format was chosen to represent the created network. Edge list is a presentation format where the vertices and edges between them are presented as pairs. In Listing 3 is an excerpt of the resulting edge list - for example (1, 2) is one pair, 1 and 2 are vertices with an edge between them. In our case, also the order of the pair is important as the direction of the edge will be available using this approach. Therefore, in the previous example the edge between 1 and 2 is directed towards the latter that is 2 vertex. Furthermore, for every pair the type of both vertices is added to the dataset and also the timestamp when both of these vertices emerged. Timestamp provides the ability to observe the graph with a temporal dimension.

Listing 3: Excerpt of the edge list.

1	2	TWEET	TWEET	1354620912	1354620151
3	4	TWEET	TWEET	1354621999	1354621671
5	6	TWEET	TWEET	1354622899	1354622829
7	8	TWEET	TWEET	1354626468	1354568981
9	10	TWEET	TWEET	1354634525	0
10	15396	TWEET	ARTICLE	1354634258	1354634256

Main task concerning the graph construction is to combine all of the outputs of the first and second phase. This means that the relations already found in previous phases can be transferred to the edge list without any major processing. All of the channels and entities have different identifier namespaces which may have collisions among them. Therefore, after the two phases the unique identifier of a resource is composed of the identifier of the resource in the channel namespace and the type of the resource. For the creation of the edge list the identifier namespaces were unified at this phase - during this process the initial identifier and type of the resource were stored in a database table along with the new identifier, making it possible to backtrack to the initial resource and its attributes. Accordingly, the relations that were found during the last two phases were also processed before adding them to the edge list by replacing the identifiers with new ones from the common identifier namespace.

One of the cornerstones in this phase was to identify relations between different entities using URLs. For this, the following approach was taken, each URL found was processed by checking whether it referenced twitter.com, facebook.com or it matched the resource locator of any of the articles that were processed. All three conditions required different operations to be done on the resource locator to confirm the relation and add it to the edge list. The following pseudocode in Listing 4 gives a better insight of the algorithm.

Listing 4: Pseudocode on how edge list was constructed.

```
for url in urlList:
      if (url contains 'twitter.com'):
            identifiers = extract tweet identifier;
      elif (url contains 'facebook.com'):
            identifiers = extract Facebook post or comment identifier;
      else:
            identifiers = find matching resource locators among article URLs
      if (identifiers is not empty):
            for identifier in identifiers:
                  store relation in edge list
```

In case of Facebook and Twitter the URLs contain required information that can be parsed for finding a matching entity in the dataset. Facebook URL may contain information about the post and comment, namely the identifiers in the Facebook channel namespace may be present in the URL and therefore can be quite easily extracted. Similar solution is available for Twitter as a tweet

resource locator contains reference to the author and also the tweet identifier. For every other URL that does no include a reference to facebook.com or twitter.com a match is searched among the URLs of all the articles that were gathered in previous phases. Before carrying out the matching with article URLs the shortest reasonable identifying substring of the resource locator is extracted - removing 'http://' or 'https://' and also the fragment identifier '#' and anything following that. If one or more matches are found all of these will be stored in edge list but not before a corresponding identifier from the unified identifier namespace is found or created. The full script used for the construction of the graph is available in Appendix 5.

3.4 Data Harvesting And Graph Construction Anomaly Removal

Exploring the newly created graph revealed some abnormalities that were the result of the work done on the dataset and also the biases of the initial datasets. These biases were not dealt with initially as there was no awareness of the results of processing the data and especially the extraction of references based on URLs. Table 1 and Table 2 indicate that the extraction and construction of the graph created a network that is somewhat biased. For this, there were various reasons that are described in more detail in Chapter 4.4. Because of this, the dataset was cleaned before any analysis was done.

Type of Vertex	Count of Vertices	% of Total
Twitter tweet	17 366	0,12%
Article	1 328 557	9,3%
Article comment	8 975 049	62,83%
Facebook comment	437 373	3,06%
Facebook like	3 393 772	23,76%
Facebook post	133 173	0,93%
All	14 285 290	100%

Table 1: Vertex distribution by type and share before data cleaning.

In Table 1 it can be seen that article comments constitute 62,83% of all the vertices, this is quite a lot and is probably of concern regarding that the relations among article comments themselves are not that interesting when studying diffusion patterns in social media and traditional media. There is also the issue of where to classify article comments - being created by people who declare their opinion rather than as a result of their work description it could be considered to be part of social media. Given that, this type of classification has not been done we here regard these as part of traditional media - in regard to the fact, that these are tightly related to articles.

Facebook likes are second by the number of vertices with a share of 23,76%. The high number of nodes of these two types of entities is due to the fact that this type of nodes have explicit relations to other entities foremost available in the initial dataset and were not discovered using URLs references. Although the article comments certainly may have URL references in their content, Facebook likes do not have any content - only their relation to a Facebook post or Facebook comment. This is also confirmed by the data in Table 2 where it can be seen that all the Facebook like vertices have only relation to Facebook posts - with the number of relations being the same as the number

of vertices. As opposed to previous entities, articles do not have any references outwards that were explicitly available in the initial dataset - of the 9,3% of the total number of nodes at least some part had an URL reference to other entities, considering that article comments relations to articles were explicitly available. The count of vertices representing Facebook comments is quite not that high - only 3,06%. Also the number of Facebook posts and Twitter tweets is unexpectedly small. The latter is probably due to the fact that the Twitter initial dataset was relatively small.

	Twitter tweet	**Article**	**Article comment**	**Facebook post**	**Facebook comment**
Twitter tweet	5025	7759		1031	
Article	144 716	55 699 514		237 462	29
Article comment	38	10 339 671	941 877	650	
Facebook post	16	2873		7435	
Facebook comment	4	49		437 491	
Facebook like				3 393 772	

Table 2: Edge distribution by source and destination vertex type before data cleaning. (Empty cell denotes value 0)

In Table 2 the distribution of edges by source and destination vertex type is presented before data cleaning. The first column shows the type of the source vertex and first row of the table represents the destination vertex of the edge. Here it can also be seen that the article comment and article relation is quite dominant, as is the Facebook like and Facebook post relation. Articles referencing other articles is the most prevalent situation in this network, with more than 55 mln edges between articles themselves. The article comment relation to articles is expected as all of the article comments have explicit relation to one news article. The article comment to article comment relations are also explicit as some news media sites enable users to comment on comments and from this another hierarchy follows. The number of Facebook like relations to Facebook post is the same as the number of vertices of type Facebook like - this shows again that the Facebook likes are the leaves in this graph.

One anomaly that was not so obvious at the beginning because there was no overview of the network - was that there were a lot of duplicate edges. This was the result of a small error in the process of constructing the edge list - but something that is easy to fix. These duplicate entries would have evidently magnified the biases in the network - therefore they were removed, in total there were 39 891 683 duplicate edges. Removing those duplicates made the graph more manageable with the tools R [2] and SUBDUE [19], used for studying the graph.

The number of article comments was another concern as it was known beforehand that some news media sites have reindexed their comments and this would result in duplicates. In Figure 3 the number of article comments are plotted, from this figure it can be seen that there are two articles which have really high number of comments - considering this as a inconsistency these comments were removed from the dataset. In total, the number is quite small but this is still discrepancy in the dataset, altogether 1855 edges were removed as a result. This plot also gives a useful insight on

the distribution of article comments - biggest portion of the articles have between 0-150 comments although the mean itself was near to 19 comments per article. The articles that have more than 200 comments are the ones that may create relevant information diffusion cascades - as it is more likely that these articles are part of some of the biggest information diffusion cascades.

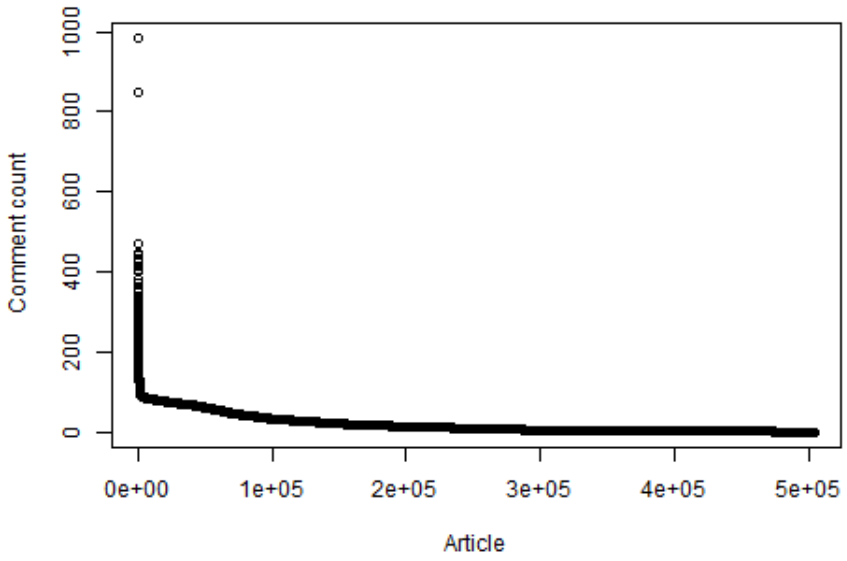


Figure 3: Article comment count.

As well few issues arose with the way the articles and their references to other entities were discovered. As articles and their references were extracted from HTML web page dumps this also included the linkages of the site that published the article. As a result there were self-loops in the network - in total 46022 self-loops were cleared. To remove other extra relations that represented web page links, a more advanced approach was taken - we looked through every article to article relation and if the edge was between articles of the same source (e.g. postimees.ee) then we looked at the timestamp. If the time difference between two articles was below some threshold, this relation was removed. Based on studying the news media sites the threshold was set to 24 hours - this is the time when most of the content of the web page can be expected to have changed. During the network processing, 8 459 797 edges were discovered that fell into this category and as a result were removed from the network.

3.5 Network Overview

For the analysis of the network multiple tools were used. As the network was quite big many tools were tested before deciding to choose one that was capable of handling this amount of data - therefore, some of the initial approaches had to be rethought. Main tool used for calculating network

metrics was R project software for statistical computing [2] and the iGraph package [7]. A lot of the information about the network was found using simple scripts written in Python - in many cases it quickly became obvious that using R for data cleaning or extracting some information was too time consuming and therefore results were achieved more efficiently using Python.

The network that was extracted from the initial dataset represents a topology of the information relations that are released in traditional online news media and also social media. Nodes in the network are of 6 different types: Twitter tweets, articles, article comments, Facebook posts, Facebook comments and Facebook likes. Article and article comment here represent the news media domain and other nodes can be considered to be representing social media. Figure 4 shows 5 examples of different subgraphs that are included in network. These subgraphs represent one information diffusion cascade - for example subgraph (a) in Figure 4 consists of nodes that reference the source article in online news media about one Estonian skier discharged of accusations of using doping [32]. Figure 4 also includes example of a diffusion cascade contained in one channel - this is the subgraph (c) that represents Twitter tweets and the retweet mechanism.

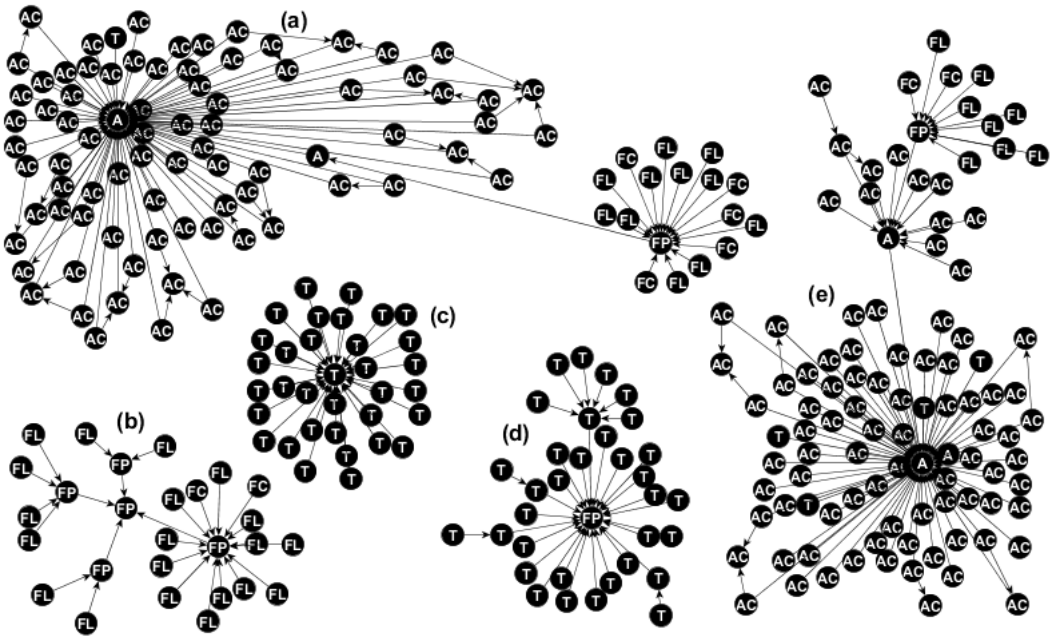


Figure 4: Excerpt of the final network with 5 example subgraphs. Edge is pointed towards the source of information. (A - Article, AC - Article comment, FC - Facebook comment, FL - Facebook like, FP - Facebook post, T - Twitter tweet)

The final network that was accumulated from various datasets consists of 14 005 945 vertices and 26 682 783 edges. In Table 3 the overview of the distribution of the vertices by their type is given. Comparing current distribution to the situation before cleaning the dataset it can be seen that the shares of different types of nodes has not changed considerably. Article comments still hold the biggest share with a 1% change being more than 64%. Facebook likes are second in order constituting a share of 24,23%. Comparing the number of vertices to the situation before data cleaning

it is visible that only article number has changed considerably - 278 688 vertices were removed. This also had a small effect on article comments with just 657 vertices removed. The proportion of the vertices with the lowest shares like Facebook comments, Twitter tweets and Facebook posts improved although not significantly. Neither has noticeably changed the division between social media and traditional media domain when using the same categorizing as described before - 71,58% of vertices belonging to traditional media and 28,42% constituting for the social media. Given that article comments constitute the biggest share of nodes in this network and therefore also for traditional media and Facebook likes make up most of the nodes for social - when removing these entities altogether the distribution between social and traditional media would stay similar, with 64% for traditional media and articles alone and the other 36% would be made up of social media entities. One thing to be noted here as well is that compared to the other entities Facebook likes differ by regard that they do not have any content created by users. It is a way of showing ones favor or sympathy towards some content created in Facebook and also revealing that certain information has propagated to people.

Type of Vertex	**Count of Vertices**	**% of Total**
Twitter tweet	17 366	0,12%
Facebook like	3 393 772	24,23%
Facebook comment	437 373	3,12%
Facebook post	133 173	0,95%
Article comment	8 974 392	64,08%
Article	1 049 869	7,5%
All	14 005 945	100%

Table 3: Vertices distribution by type and share.

The edges of the network are directed - they point towards the information node that is referenced by the starting node, therefore the direction does not represent the flow of information rather than the source of the information. Additionally, a large part of the vertices have a temporal dimension that refers to the date and time when the edge became available and therefore also the time when the tail of the edge was created. With the temporal dimension the texture of the information diffusion in this network will be described in Chapter 4.3. Among other characteristics of the information diffusion network is that the graph is not connected - this results in that the subcomponents of the graph represent different diffusion cascades.

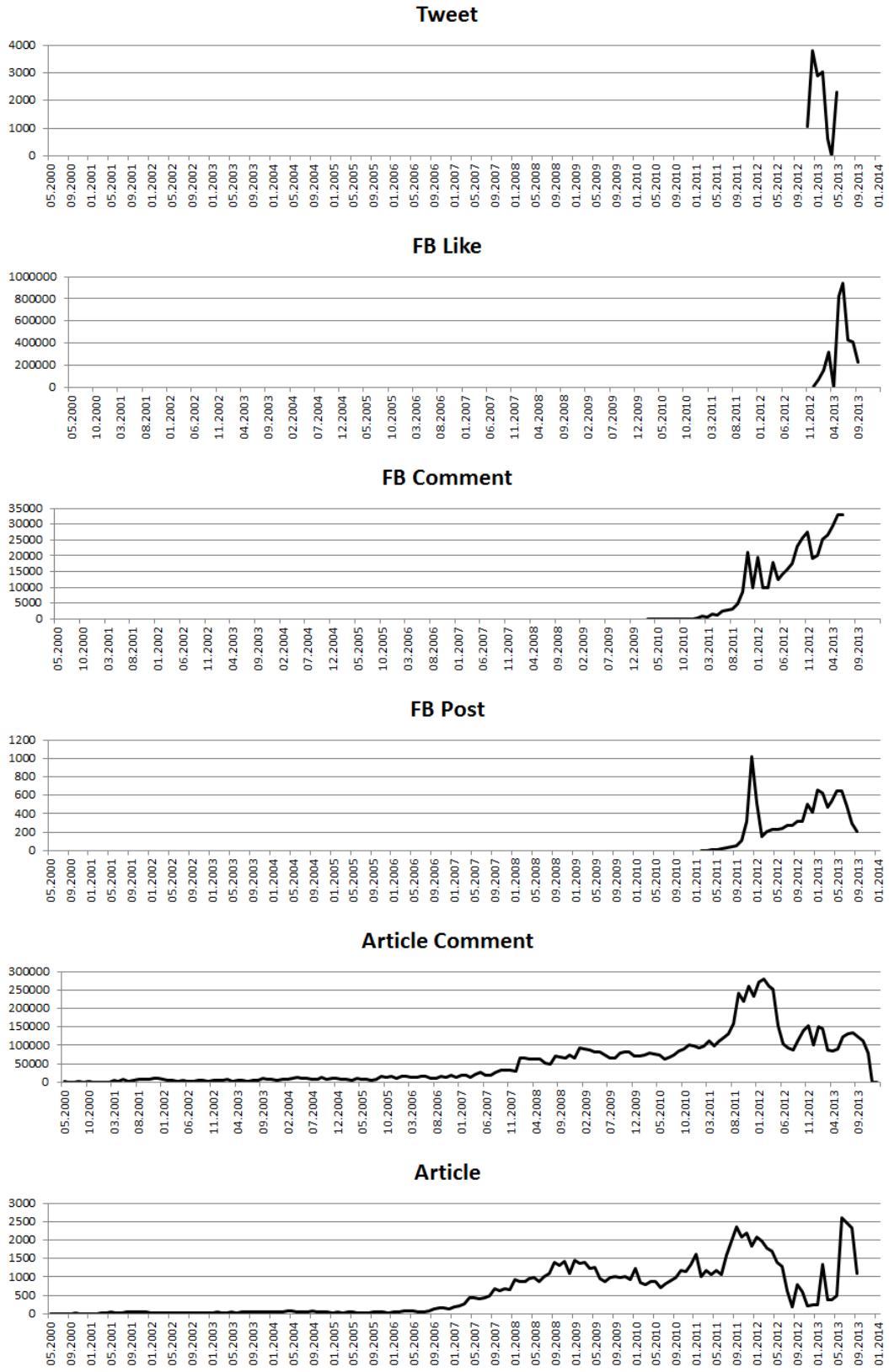


Figure 5: Vertex number timeline by vertex type

In Figure 5 the number of vertices created over different time periods is presented. Different type of vertices are from different time periods - for example there are article and article comments over 12 years while tweets and Facebook likes are from a time span of 7 and 9 months. Although it is visible that there is a common period for all the different vertices from December 2012 till May 2013. The chart representing the timeline of articles also indicates the rise of the importance of online media during the last 6-7 years.

	Twitter tweet	**Article**	**Article comment**	**Facebook post**	**Facebook comment**
Twitter tweet	5025	7759		1031	
Article	135 801	12 127 014		161 295	
Article comment	38	9 538 272	864 342	634	
Facebook post	16	2861		7352	
Facebook comment	4	49		437 491	
Facebook like				3 393 772	

Table 4: Edge distribution by source and destination vertex type. (Empty cell denotes value 0)

Graph edge distribution based on source and destination node in this network is another metric that in large scale gives a good overview and description of the information diffusion process in social and traditional media. Given that, there are 6 different types of vertices, therefore altogether there could be 36 different types of edges but in this dataset for example an edge between Facebook like and an article is impossible. In Table 4 the distribution of edges by source and destination vertex type is presented. The first column shows the type of the source vertex and first row of the table represents the destination vertex of the edge.

In Table 4 the changes compared to the situation before data cleaning are more visible - the number of edges has reduced from 71 219 412 to 26 682 783. Most of the edges that were removed were article to article edges, but still this relation is the most dominant one. While the removed article to article edges were mostly duplicates one of the biases was also reduced by removing edges that did not represent diffusion process rather website linkage. Table 4 first two columns and the third one with the exception of the last row (values with underscore) belong to the core of this information diffusion network. These are the edges that most likely will interconnect two domains and have vital role in information diffusion cascades. Quite interesting phenomenon is the relatively high number of articles referencing Twitter tweets and Facebook posts. Somewhat unusual is that in the initial network there were 29 references from news media articles to Facebook comments - looking into the dataset, it was discovered that it was reference to one specific Facebook comment by all the articles and as a result of data cleaning these references were removed. Therefore, what is in common with Facebook likes - comments as well are only leaves and do not play central position in the diffusion process. Inspecting Facebook post references to Twitter tweets where no inconsistencies were found, as was the case with other edges that constitute a quite small amount.

3.6 Metrics of the Network

One group of the graph metrics analysed are the centrality measures that describe many aspects of information diffusion in a social network or different hybrid networks. "Centrality metrics are point-measures on the network, allowing the measurement of the power and infuence of individuals in a social network", this is how the centrality metrics are described in diffusion study by Steinkirch [42]. In the social and news media hybrid network these measures rather than describing the importance and influence of individuals - they describe the relevance of nodes in the flow of information diffusion. Moschalova and Nanopoulus [28] have studied the network centrality measures with regard to finding the best seed to propagate some information rapidly. Their study concentrates on four most used centrality scores: degree, betweenness, closeness and eigenvector. Although the intention of their study is to find which of these scores maximizes the spread of information, still for example closeness centrality scores describes the size of the information diffusion and PageRank being a variant of eigenvector centrality characterizes the connectedness of the nodes in one information diffusion cascade. PageRank will give an overview of the significance of specific types of nodes in the network.

In Table 5 the results of calculating the betweenness centrality measure and PageRank are available by the type of the vertice. The betweenness centrality here helps to quantify the importance of different types of nodes in the diffusion processes. Having identified the influence of the nodes provides insight on how the results correspond to the number of edges between different entities in Table 4 and whether the quantity of edges is proportional to the importance of the types of starting and ending nodes. The results in Table 5 (with cutoff value set to 5 - which is the maximum path length considered when betweenness was calculated) show that for the betweenness centrality measure the average number of shortest path that pass through articles is extremely high. This quite well quantifies the importance of articles in the network but this high betweenness score is also somewhat a result of the high number of article comment and article relations. For Facebook like and comment betweenness score was not calculated as previous metrics have shown that these two entities are leaves of the graph and therefore no shortest paths pass through them. Due to the big ratio of vertices of article type the overall betweenness score is quite high as well. Considering other types of vertices, Facebook posts also tend to have more central role in the information diffusion cascades. It could very well be that articles and Facebook posts are at the periphery of the social media domain and media domain in the information diffusion cascades connecting these two domains - this can be evaluated when the topological patterns are studied. The betweenness score for Twitter tweets and article comments is not that significant but will definitely have some interesting implications to some of the information diffusion cascades.

	Twitter tweet	**Article**	**Article comment**	**FB comment**	**FB like**	**FB post**	**All**
Betweenness	3,297	30300	0,46	NA	NA	663	2278
PageRank	$3{,}01\ 10^{-7}$	$3{,}38\ 10^{-7}$	$3{,}97\ 10^{-8}$	$3{,}82\ 10^{-8}$	$3{,}82\ 10^{-8}$	$1{,}03\ 10^{-6}$	$7{,}1\ 10^{-8}$

Table 5: Mean vertex betweenness and PageRank scores by vertice type.

PageRank, another centrality measure scores in Table 5 (with damping factor 0,85) are extremely low for every type of node and also for the whole graph. Only Facebook post score is one magnitude higher than article or Twitter tweet score - this agains confirms the significance of Facebook posts and their central role in the information diffusion processes. Article comments, Facebook comments and Facebook likes scores are another magnitude lower. The metrics calculated so far have indicated some characteristics of the network - articles and Facebook posts have far more vital role in the way that news media and social media content flows through the network.

Another graph metric that is interesting in the context of this type of network is the distribution of in- and out-degree of the vertices. Leskovec et al. in their study [25] looked into the distribution of in- and out-degree and found that for blog network the assumption of the amount of in- and out-degree being balanced for popular blogs was incorrect. In this network, we ought to see if there are any discrepancies in the distribution of node degrees based on the type of node - whether articles tend to be referenced more and is social media mainly responsible for the propagation of the information. In Table 6, the average in- and out-degree has been calculated for every type of vertice and the whole graph. The results correspond to the different types of edges brought out in Table 4 - Facebook comments and like average out-degree is 1 showing that every entity of this type is connected to Facebook post by an edge. The out-degree of Twitter tweets is somewhat interesting - the 0,8 of Twitter tweet implies that almost every tweet reference at least one other entity - this can be by the means of a retweet or a URL reference. Furthermore, the mean in-degree of a Twitter tweet 8,11 reveals that a tweet in this network on average has more than 8 references to it - visible from Table 4 that biggest share of these are references from articles. 0,08 average out degree of Facebook posts can be expected to mostly be referencing articles.

The number of edges from articles reveals that on average there are almost 12 other entities referenced. This high out-degree could indicate some deficienies in the data cleaning concerning not relevant references that were included in the HTML page - this is considered more closely in Chapter 4.4.

	Twitter tweet	**Article**	**Article comment**	**FB comment**	**FB like**	**FB post**	**All**
In-degree	8,11	20,65	0,1	0	0	30,05	1,9
Out-degree	0,8	11,83	1,16	1	1	0,08	1,9

Table 6: Average vertex in and out degree by vertice type.

The degree distribution without regard to edge direction is following power-law with exponent ~2.7 as exhibited in Figure 6 - this indicates that this is a scale-free network. Although Figure 6 indicates some obscure situations with the degree distribution - some occasional rises in the frequency at the right hand side. A common feature of scale-free networks is that there exists a small number of nodes with high degree, these nodes are called the hubs [5]. The role of these hubs as influencers has been also described by Janssen in his study on simulating market dynamics and researching the role of hubs in social network and market decisions [18]. In this network there are 12535 hubs that have degree above 500.

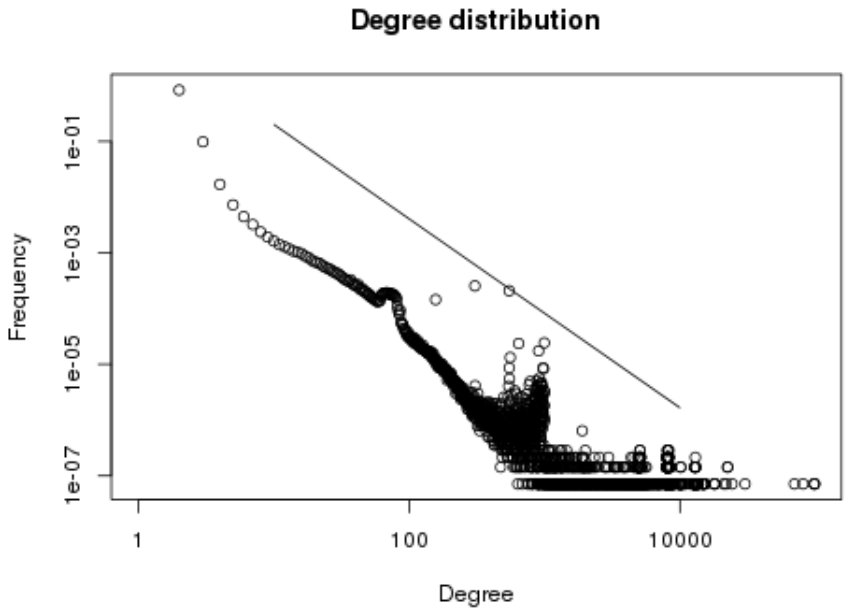


Figure 6: Vertex degree distribution fit power law.

Assortativity is another measure that is closely connected to in and out degree - in network, this is defined as the phenomena whereby nodes that are similar in some manner tend to be connected. The similarity may be in that the nodes with high degree are preferentially associated to other nodes with high degree studied by Newman [31] and this was found to be common for social networks. In this network this is not the case - the assortativity coefficient being -0,031 indicating no assortativity by the degree. This also reveals that the cores of the diffusion cascades therefore tend not to be dense.

The number of connected components allows to have a better understanding of the structure of the network. The existence of cliques with a large number of nodes in this network is highly unlikely unless these are some anomalies that are present in the network. Connected components will give a useful insight on how big the information cascades are and as well the significance of those structures in diffusion process. Figure 7 presents the distribution of the connected component sizes in three separate plots displaying three groups of different size - components of size 1-100, of size 100-1000 and components bigger than 1000. As it is visible in the plot of components of sizes 1-100, the biggest share of them are of size less than 5 - this reveals that most of the information cascades are not noteworthy and tend not to get a lot of attention by other participants of the network. Besides the first plot contains the sizes of 98% of all the components - therefore, a huge portion of the components are of size less than 100 nodes. Probably the most interesting cascades are those which are bigger in size - in the plot of components with size greater than 1000 there are few components of size 10 000 and above, one component that consists of more than 55 000 nodes.

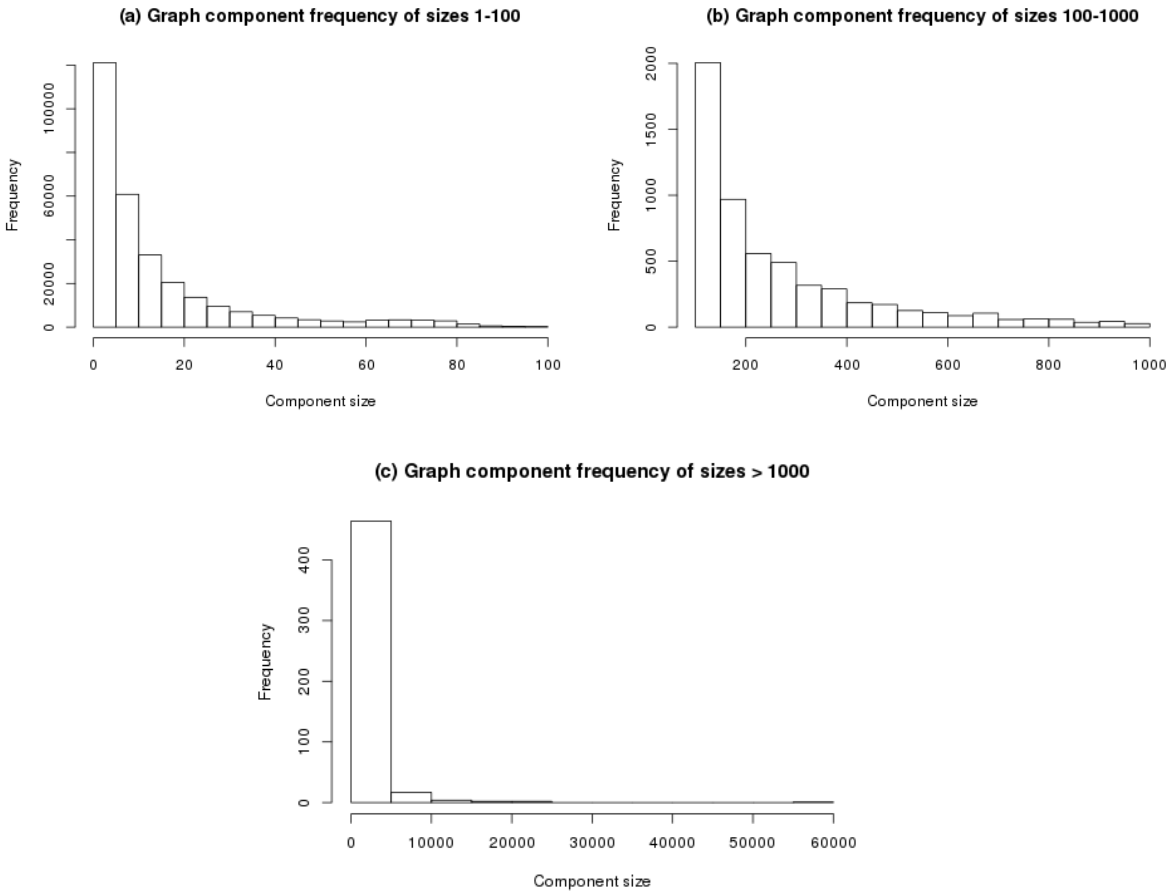


Figure 7: (a), (b), (c) Distribution of the sizes of the connected components.

4 Information Diffusion Pattern Analysis

The diffusion patterns that are relevant in this multidomain network that is composed of vertices of different types are explored next. The patterns that are present in the graph are studied from two perspectives - first, the temporal perspective gives an overview about the characteristics of how information diffuses over time and what are the diffusion cascades' dominant behaviours. Secondly, the topological patterns are investigated - this gives an overview on what role each type of entity plays in the diffusion processes.

4.1 Temporal Overview

One subgraph in this network represents an information diffusion cascade and for this all of the subgraphs in the network are extracted - in total, there are over 300 000 subgraphs. For every subgraph time difference between every node and the root is calculated. In Figure 8, the total number of references that emerged during the first year from the creation of the initial information is available. Here it is visible that huge part of the references emerge during the first day after the release of the information - also significant number of referencing entities are created during the first week. The first days after the first week of the diffusion cascades here again indicate high amounts of information diffusion taking place and after this there is a decay until the 50th day after the initial release - then, a slower decay follows with some interesting spikes just after the 50th day, also near the 120th day and later on near 300th and 330th day. These maybe a result of some information becoming relevant again at a later point in time. Otherwise, it is vivid that almost every information cascade is really active during the first day and also the first two week after the release of source information.

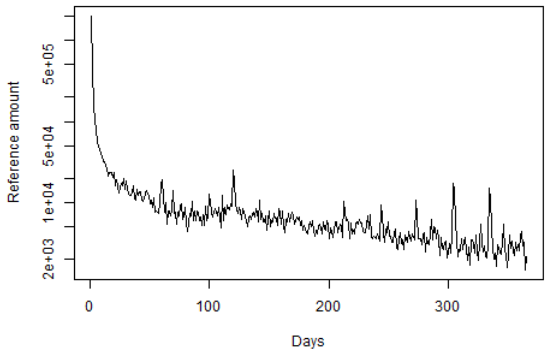


Figure 8: References created during one year by day (y-axis in log scale).

Based on the significant activity during the first two weeks, this period was extracted and looked into with a more exact time interval of hours. Figure 9 shows the number of references emerging every hour during the first two weeks. Here, the view of the diffusion process is similar when the time unit was day as in Figure 8 - the first few hours are notable in the number of diffusion events that take place. After the first 12 hours, the number of diffusion events taking place starts to level

off - this implies that during the first day after some info is released it diffuses quite rapidly and it declines at the end of the day. The smaller peaks after this show that the next day the information diffusion process continues with a slower rate - the peaks continue to occur with a daily pattern.

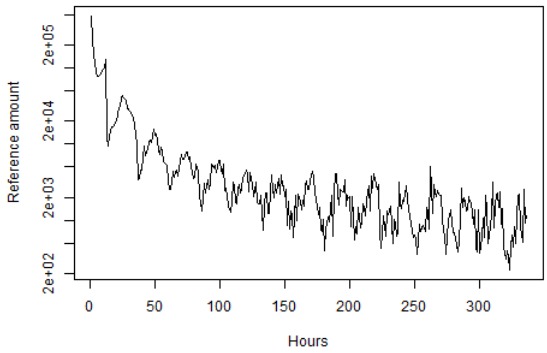


Figure 9: References created during first two weeks (y-axis in log scale).

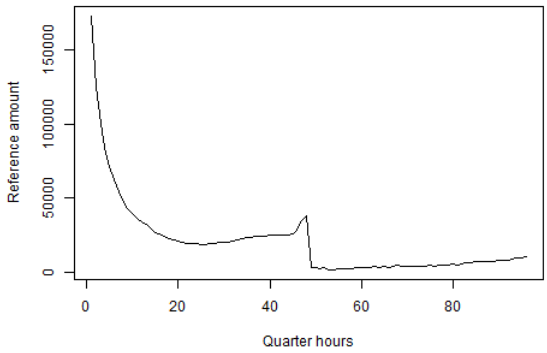
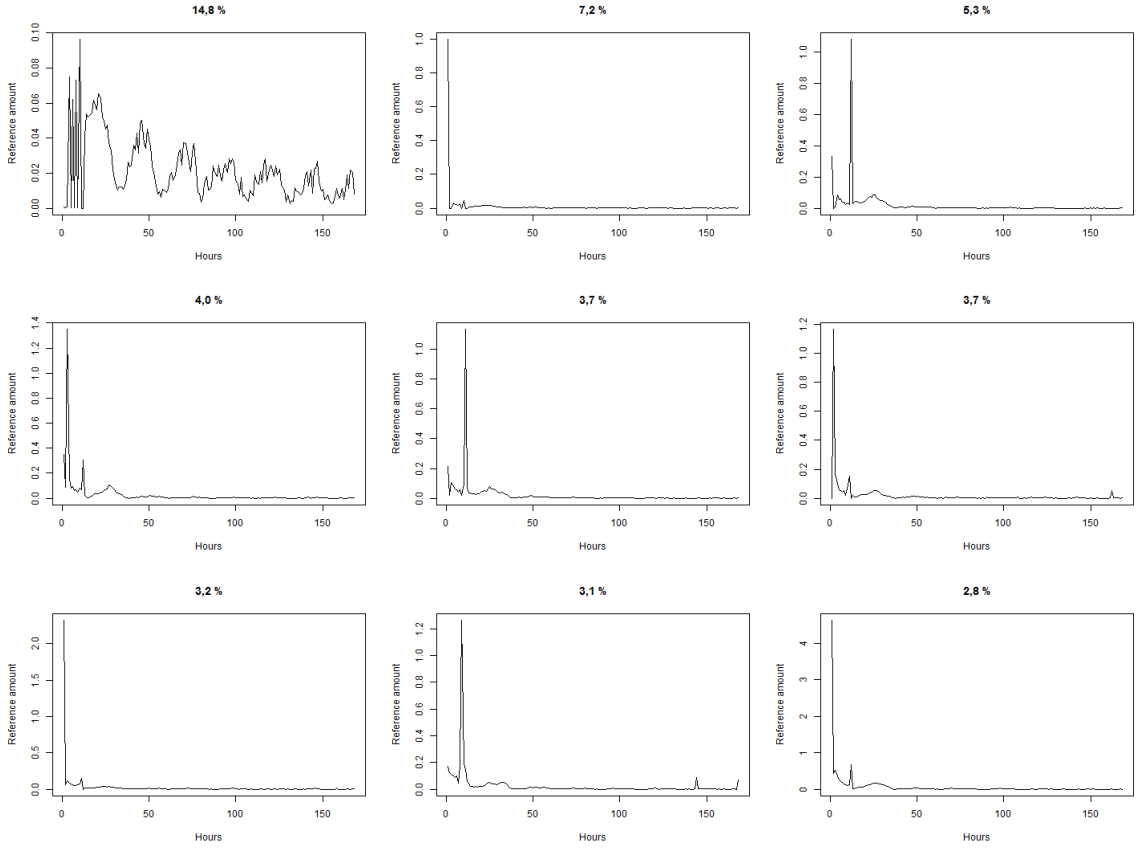


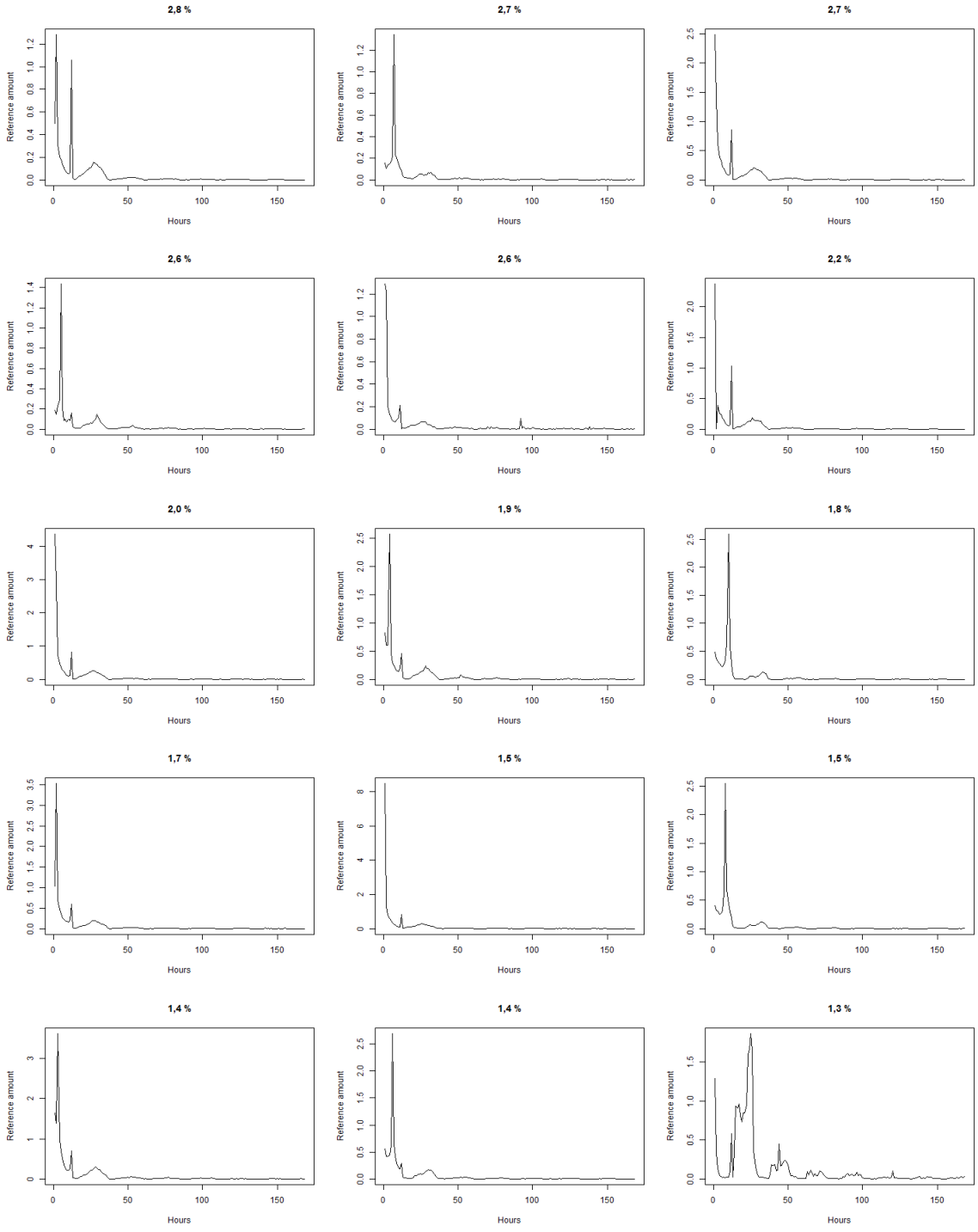
Figure 10: References created during the first 24 hours by 15 minute unit of time.

Figure 10 shows how the information diffuses during the first day with a time unit of 15 minutes. Here again first 2-3 hours are the most active period - this means that the pace in which information diffuses in this network of social and traditional media is really quick. The peak and a sudden drop that occurs after nearly 12 hours has passed from the initial release of the information is an interesting phenomenon - one reasoning for this could be the daily routine of people and this drop falls just before the night. Next paragraph can more precisely explain the backgrounds of this obscurity.

4.2 Temporal Patterns

The subgraphs were extracted and for every subgraph, a vector of the number of referencing events during the first 14 days by hour from the creation of the initial information were created - events that did not belong to this period were excluded. Next, the vectors were clustered using k-means clustering method. This is a somewhat similar to the approach taken in an article about network evolution in Skype by Kikas et al. [20] studying the contact addition patterns. The activity in the information diffusion cascades decays significantly after the first 14 days. The 14 day time span was chosen based on the data available in Figure 8 where it can be seen that after 14 days the total number of new references is less than 25 000, given that there are over 317 000 subgraphs this indicates that the possible patterns that emerge later will not be significant also based on Figure 9 where there is intense activity during this period. Some of the subgraphs were excluded as none of the nodes had a timestamp - therefore, these did not expose any important value for the pattern extraction. Resulting clusters describe the most common cases and these are analysed next to identify other common factors that are inherent to the cascades that can be found in this type of network.




2,8 %
2,7 %
2,7 %
2,6 %
2,6 %
2,2 %
2,0 %
1,9 %
1,8 %
1,7 %
1,5 %
1,5 %
1,4 %
1,4 %
1,3 %
Reference amount
Hours

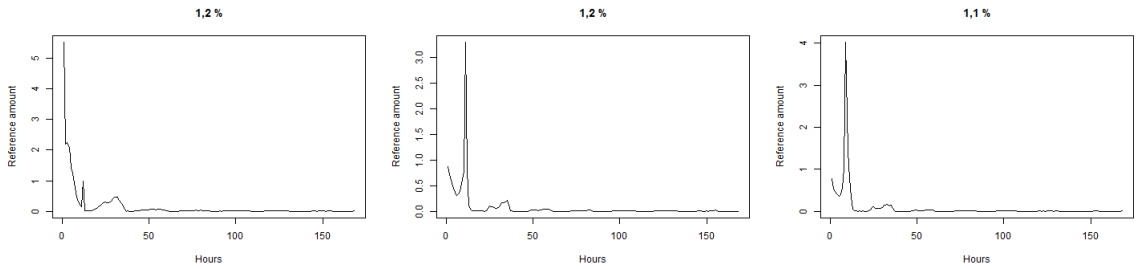


Figure 11: Most common information diffusion temporal patterns.

The results of clustering are represented in Figure 11 where the most significant clusters are available - every cluster that had more than 1% of time series belonging to that cluster. The 0th time period in these figures represents the period of time from the release of the information till one hour having passed. The common aspect visible in all the clusters is the length of the initial latency - which shows that the peak of the time when information diffusion actively takes place is quite early during the first 24 hours or in most cases even during the first 12 hours of the initial information release. As well the high importance of the first hour is visible as this is the time when a lot of diffusion events take place - this is similar to the results found in [22] by Kwak et al. studying the Twitter network. Another pattern that is present is the daily surges - especially in the first cluster and this also obvious in Figure 9 although as the number of references emerging during later periods is quite low, this is still a common factor. In the first cluster presented in Figure 11 the calculated average difference between the peaks was 22,76 hours - which confirms the daily pattern. This pattern would have probably not emerged if this network was not geographically inclined towards Estonia due to the initial dataset - in global dataset the daily patterns of this setting would have not emerged.

Most of the clusters visible have very low number of information diffusion events taking place showing that the usual size of the diffusion cascades is not that big. The average number of references added at every hour is quite slow revealed by the scales of the y-axis - only few of the clusters have more than 4 diffusion events taking place at some hour during the first two weeks.

Among the common patterns there is not a single cluster where the highest peak occurs after more than 50 hours having passed - this indicates that the pace at which the information diffuses is quite high and also the relevance of the information is immediately estimated by other participants of the diffusion process. Although none of the clusters indicate any kind of viral diffusion patterns given that the reference number is not as high as could be expected in these cases. There are 116 subgraphs that have more than 2000 nodes which is about 0,04% of all the subgraphs which is also visible in Figure 7 - as this number is quite small, the viral diffusion pattern is not significant in this network. The 2000 node criteria chosen here is based on the study by Goel et al. [14] about the structural virality of online diffusion where the news channel is shown to have a noticeable structural virality score with cascades of size 2000 and higher.

One thing that seems to be prevailing in many of the biggest clusters is the existence of a later

burst - this burst takes place 10-12 hours after the release of information. In few of the cases, this second peak is also the highest - indicating that there exists some delay when the information reaches other entities. It could be explained by many factors - but one reasoning for this could be that the creation date of the initial information can differ from the release date, which is not available for this network. Other case is the small rise in diffusion events approximately 32 hours after the initial release of the information. The cause of this could for example be that online news media tends to keep some topic active and bring it back after day or two to get more visitors. This situation tends to occur just before the number of new references created becomes really low and the calming of the diffusion process.

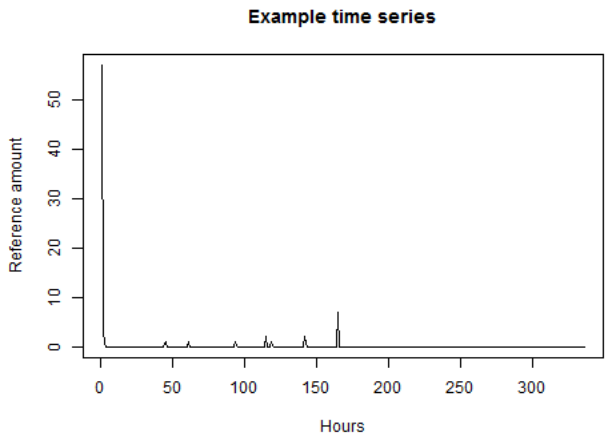


Figure 12: Example temporal pattern of the release of news in Estonian online media about Estonian sportsman Andrus Veerpalu being found not guilty for doping by Court of Arbitration for Sport.

Figure 12 shows one example of a temporal pattern available in the network. This is the same subgraph for which the topological pattern is available in Figure 21 - it is visible that this pattern is in compliance with the patterns found in clusters, during the first hour large part of the diffusion takes place and subsequent events tend not to create any diffusion bursts. It is quite difficult to place this temporal pattern into any of the clusters found before - especially due to the extremely high burst at the first hour compared to the peaks in the clusters. Taking this pattern into pieces one by one: the peak at first hour is mainly composed of the comments written to the article; later on at about 120 hours later a new article referencing the inital one is released and a Facebook post follows immediately after; the second highest peak is the result of Facebook comments and likes. This pattern is somewhat unique as it contains events from multiple different channels - as in this network majority of the cascades tend to be confined into one of the channels - for example article and article comments or Facebook posts, comments and likes.

4.3 Topological Patterns

While temporal patterns give an insight into when information diffuses, topological patterns are studied here with the intention of finding out how the information diffuses. Topological patterns are studied from the perspective of different network motifs found to be present in the graph - for this, the subgraphs containing less than 3 connected vertices were removed as these small motifs do not carry any sort of interesting information about diffusion process. Topological patterns were extracted using frequent subgraph mining techniques - SUBDUE algorithm [19] was used to extract frequent patterns in the subgraphs. Frequent cascades have also been studied by Leskovec et al. [25] in blog networks. The approach taken in their study differs quite a lot from the approach taken here - first they extract only cascades, not patterns and they are restricted to one domain that is the blog domain - in this thesis the patterns can include nodes of different domains available in the network.

To discover topological patterns, current subgraph and motif discovery methods were explored. As identifiying different patterns is computationally expensive, some of the algorithms described in research papers were not suitable for graph of this type and size. SUBDUE which is a compression-based algorithm for frequent pattern discovery in graphs was used to extract topological patterns - developed by Ketkar et al. [19]. SUBDUE is based on the Minimum Description Length principle - beginning with each different type of vertex in graph substructures are created by expanding them - substructures that give best results in compressing the whole graph will be used in the next step until no quality substructures are left. For the presentation of subgraphs and patterns yEd Graph Editor [3] was used. Frequent patterns describe the most common situations on how information diffuses between different domains or being contained only in one domain.

The subgraphs that had only 2 vertices connected were removed - in total 558 776 vertices were of this kind. For the input, the full graph was divided into chunks that consisted of about 30 000 sub-components each as this was the only way results could be retrieved with meaningful time. Frequent patterns of sizes 3-10 were extracted - this required in total 64 iterations to be run. The number of frequent patterns to be extracted by SUBDUE software was set to 15 as initial experiments with the data revealed that it is highly unlikely that more than 10 patterns will be discovered. After the 64 iterations of running the algorithm were carried out the results of different chunks were combined such that the most common patterns of sizes 3-10 became available for analysis.

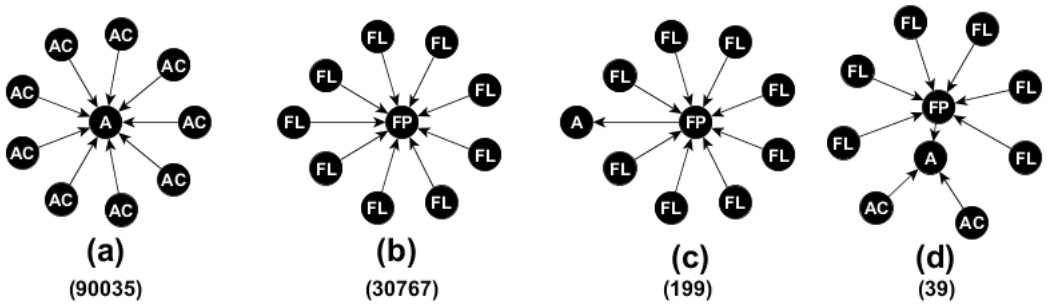


Figure 13: 4 most frequent subgraph patterns of size 10 (Number below shows the number of patterns discovered).

Figure 13 exhibits frequent subgraphs of size 10 in the network. Two types of patterns (a) and (b) are by far the most common in these networks - article comments referencing an article and Facebook likes referencing Facebook post. These patterns are also contained in one channel that is news media and Facebook channels and therefore do not indicate what are the patterns that affect the crossing of channels or even domains. The two other patterns (c) and (d) are slightly more interesting as these depict the information diffusing across domains. Pattern (c) shows a Facebook post referencing a news article which itself is referenced by multiple Facebook likes therefore this is information diffusion from news media domain to social media domain, specifically Facebook channel. This diffusion process signals that some information diffusing to Facebook is widely received by multiple entities in this channel as represented by the number of likes. Pattern (d) is similar to (c) but this pattern also includes the diffusion process among the news media channels as well by the article having two comments. Pattern (d) in Figure 13 shows that the diffusion takes place in one channel and also the other channel with an article and Facebook post creating a bridge between those two diffusion subcascades. There are two issues to consider here as well - first, these patterns are limited to size 10 so taking pattern (d) as an example, it can be expected that the article is referenced by more than two article comments and Facebook post can have Facebook comments as well. The other issue is that extracting patterns was limited to that no overlapping was allowed so some of the most popular patterns could have captured most of the nodes that could have been part of other patterns as well.

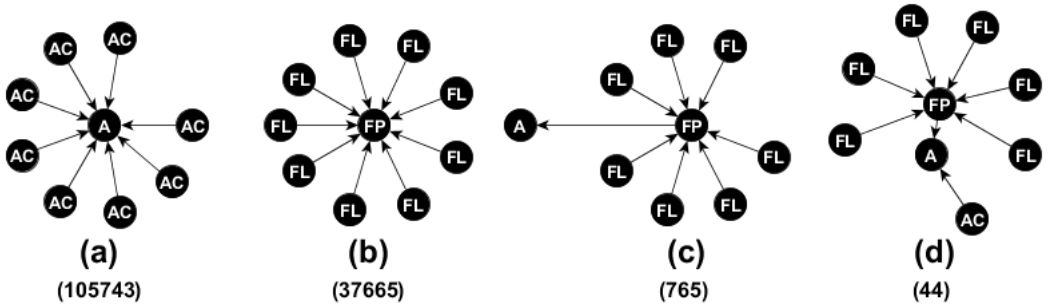


Figure 14: 4 most frequent subgraph patterns of size 9 (Number below shows the number of patterns discovered).

Figure 14 shows motifs of size 9. These patterns are almost exactly the same as those in Figure 13 only one referencing node less. The total number of pattern (a) found in the graph is another indication of the vital role news media and especially articles play in this network. Besides articles, Facebook post is another essential entity in information diffusion processes in this network - although likes do not contain any textual content, they signify that this information has reached other participants of the network.

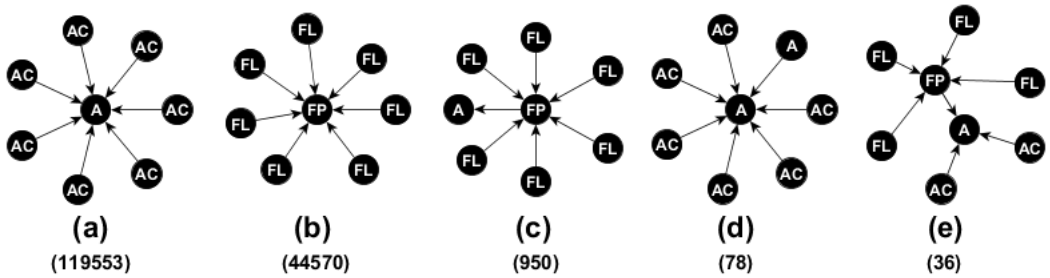


Figure 15: 5 most frequent subgraph patterns of size 8 (Number below shows the number of patterns discovered).

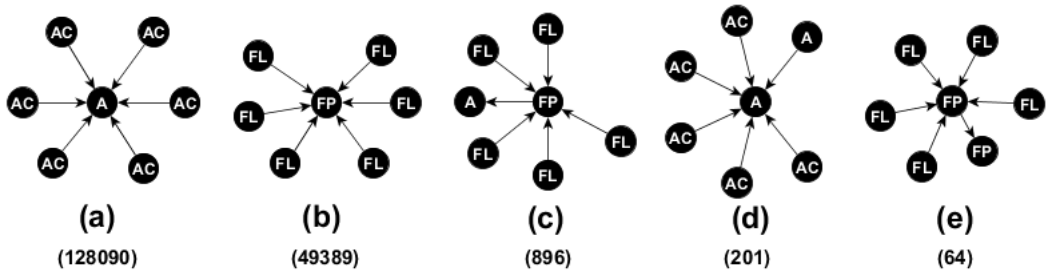


Figure 16: 5 most frequent subgraph patterns of size 7 (Number below shows the number of patterns discovered).

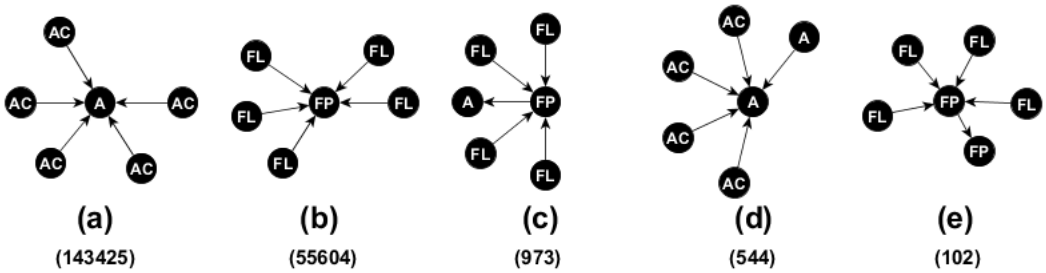


Figure 17: 5 most frequent subgraph patterns of size 6 (Number below shows the number of patterns discovered).

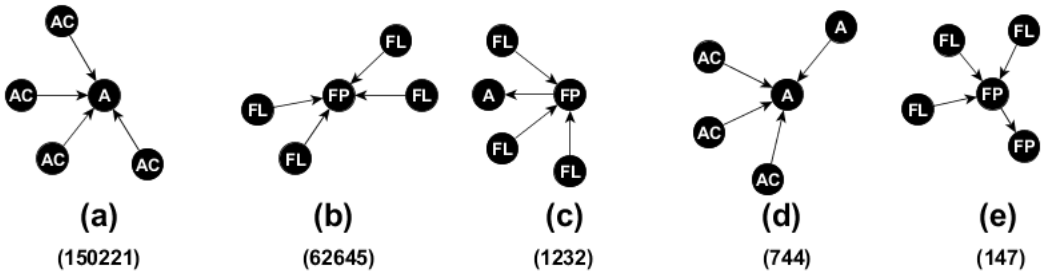


Figure 18: 5 most frequent subgraph patterns of size 5 (Number below shows the number of patterns discovered).

Figures 15, 16, 17 and 18 show most frequent patterns of sizes 8, 7, 6 and 5 correspondingly. While the patterns got smaller in size more patterns emerged that are frequent in the network. For example, pattern (d) in Figure 15 is a new pattern that is quite similar to the most common pattern with only difference being in that the central article is also referenced by one other article. Two patterns (a) and (b) that are the most common by a great margin differ from the others apart from pattern (d) as they are star-shaped and as such can be regarded as the core of the diffusion. Pattern (d) and (e) in Figure 18 indicate the flow of information between the core entities of news media and Facebook channels with the article to article diffusion and Facebook post to post diffusion - as the size of the patterns decreases, the number of the most frequent patterns increases, showing that these patterns are the most important and considering the number of vertices and edges in the network, there exists a wide variety of ways how information diffusion takes place. All of the frequent patterns this far have not included Twitter tweet nodes - this is explained by the low number of vertices of this type only 17 366 as is presented in Table 3.

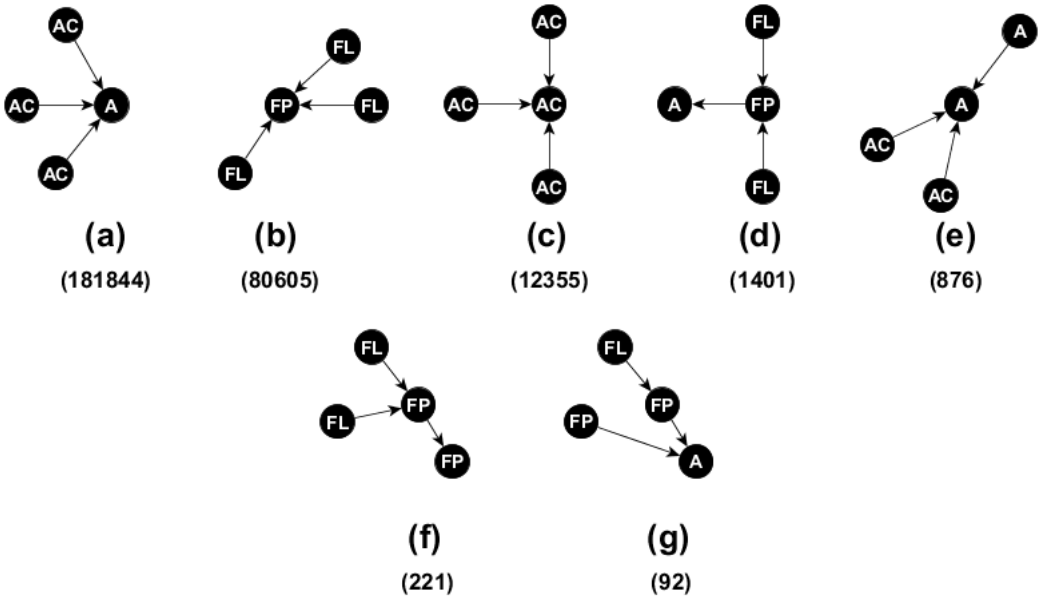


Figure 19: 7 most frequent subgaph patterns of size 4 (Number below shows the number of patterns discovered).

Figure 19 presents 7 motifs of size 4. With the lower size of the motifs there exists more relevant patterns that describe information diffusion. Pattern (c) is one that was not available with the bigger patterns and it shows the diffusion process among article comments - a situation when there is an active discussion on some comment given. As article comments exist only with relation to articles we already know that every comment in that pattern is related to the same article as well. Pattern (g) is another addition and also represents the diffusion process between social and traditional media.

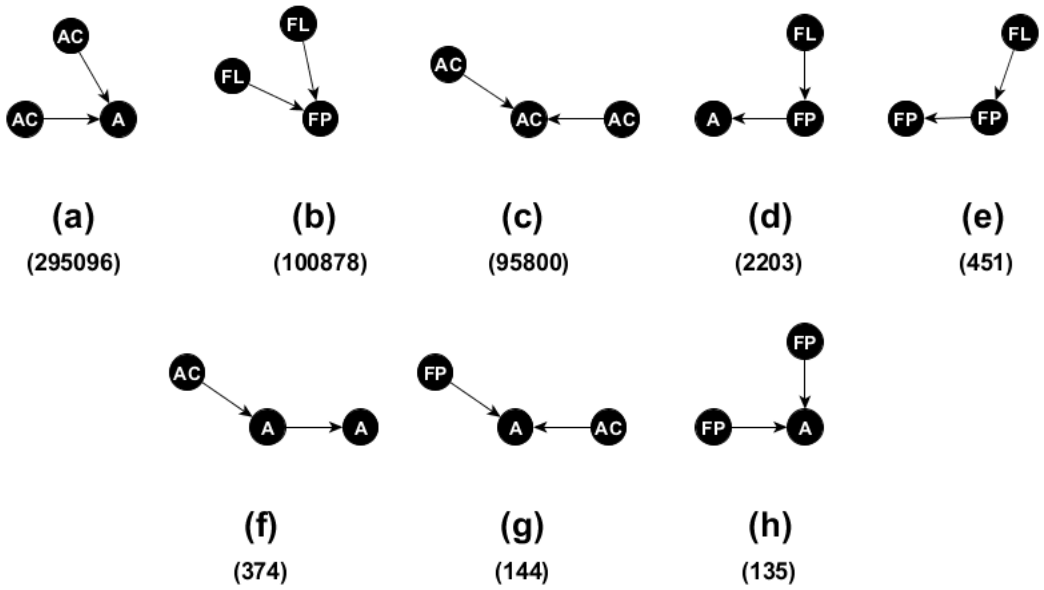


Figure 20: 8 most frequent subgraph patterns of size 3 (Number below shows the number of patterns discovered).

Figure 20 presents the foundation motifs of the network. 3 node patterns depict the core diffusion directions and processes that take place in this network. The high number of article and article comment diffusion patterns still remains the top diffusion process and confirms the importance of news media in this network. The relatively big growth in the number of pattern (c) is another confirmation of the article comments being an active place for discussion of different daily news media topics. There exists a similarity between Facebook channel and news media channels - both have a source object that initiates the diffusion process, in case of news media it is an article and for Facebook it is a post. Facebook comments and article comments are entities that describe how the information diffuses and how the discussion of the source information takes place. This indicates the central role of Facebook posts and articles in this network and that the diffusion in most cases starts with information released in a form of one of these entities. The cascades that these patterns initiate depict a star-like shape where the information propagates starting from the initiating information and through multiple nodes the information diffuses till reaching a point in time when the information has lost its relevance. One example of this is presented in Figure21.

Two of the patterns (e) and (f) in Figure 20 also present the inter-domain diffusion pattern but with the difference in that information diffusion is between core components of the network. This indicates that some new information related to the source became available and is published - creating another source of a widespread star-shaped diffusion process in the domain.

Another set of patterns are (d), (g) and (h) are different from the other foundation patterns in a sense that they incorporate diffusion accross multiple channels. In all the patterns, the diffusion direction is from news domain to Facebook domain and none of the patterns have indicated the prominence of the diffusion the other way around - this unidirectional diffusion conforms to the

findings by Kim et al. [21]. This thus shows that the information sources or the topics of discussion are mostly initiated by the news media - although there are some rare cases when something published in social media catches the attention of some news media participants.

Whole diffusion process in this network could be described as a two level diffusion process - at the first level is the diffusion that is bounded by one specific domain. Diffusion between article and article comments is one example of this kind of diffusion, in Figure 20 patterns (a), (b), (c), (e) and (f) belong to this group. Second level is the diffusion that crosses domain boundaries for example a Facebook post referencing a news article, in Figure 20 patterns (d), (g) and (h) belong to the second level.

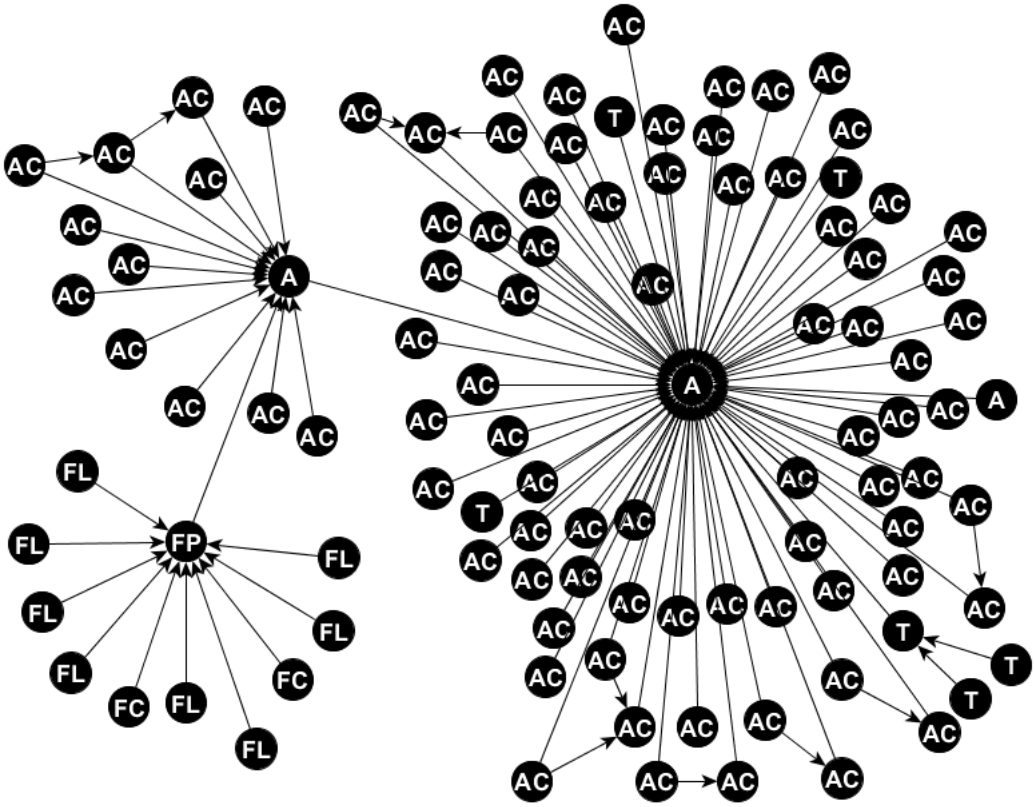


Figure 21: Example subgraph (information diffusion cascade).

Figure 21 exhibits an example subgraph that was extracted from the network. The subgraph depicts the same information diffusion cascade that was analysed from temporal perspective in Figure 12 - the source news article being about the Estonian sportsman Andrus Veerpalu being found not guilty for using doping by Court of Arbitration for Sport [32]. This subgraph includes most of the patterns in Figure 20 except (e) and (h) being a good example in showing what the full diffusion cascade includes. This example incorporates diffusion to all the channels of the network - news articles, Facebook and Twitter channels. Here, the processes discovered in studying specific patterns apply as well - for example, Facebook post and articles are part of the domain crossing

diffusion, with the addition of Twitter as well. In the borders of one channel there is an active diffusion process where the star-shaped subcascades are formed. Here, the two level diffusion model can be easily captured with the main diffusion line of article to article to Facebook post being part of the second level diffusion and the diffusion cascades starting from these nodes part of the first level diffusion.

4.4 Threats to Validity

During data extraction, construction of the network and analysis some issues were discovered that may threat the validity of the analysis in some part and therefore the conclusions based on this. One of the issues was that the URLs found in some of the inputs were not validated, which is a slight concern with news media. Initially the goal was to carry out the validation process as well but few issues arose at early phase: validating URLs by making HTTP queries was too time consuming; secondly, the risk of our servers being black-listed was too great, given that number of queries would have been really high. This means that some of the relations between entities as a result were not found.

Another note about the network is that nodes that belong to social media are slightly biased towards information that is created by people who are board members of Estonian companies - this is due to the fact that the initial data was crawled for extracting information about these people. Although, this should be a small bias as the URL references found in articles and other entities to any social media entity are not restricted by this criteria and were added to the dataset.

There are two minor issues that need to be considered as well - these issues are mainly with regard to the articles and article comments that were used as an input for creating the network. Firstly, the approach taken for extracting URLs from article HTML dumps meant that the URLs that reoccured in some news media HTML pages were extracted multiple times and in many occasions they might have not been relevant to the content of the article. For example, postimees.ee news articles are presented in a manner where there are multiple other links to traverse the web page and also related topics or even the last published news stories. Here one can argue if this kind of URL reference represents a relation between two objects or not, given that this in some cases could be thought of as representing the web page interrelations. To ignore these kind of relations first 24 hours of every article to article diffusion was validated. Second issue is a concern primarily with the online media domain dataset as sometimes the news articles are reindexed or renamed - this could have resulted in that some of the articles or article comments were processed multiple times, which also meant that there may be reoccuring nodes in the network. This type of situation is not prevalent and thus far only one case has been found - to remove these kind of situations a lot of overhead computing would have been necessary and the resulting bias is not significant.

5 Conclusions and Future Work

The focus of this thesis is to study information diffusion processes in a network that incorporates multiple networks of different domains. The network under study here was built upon information collected from different social media channels and traditional news media channels. The processes of information diffusion were under study from two different viewpoints - the temporal and topological viewpoint. These two aspects of the diffusion processes help better understand when certain events tend to take place and secondly how these information diffusion events happen.

Before studying the network could begin it had to be constructed on the basis of different initial datasets - this included web page dumps of news media, database with collected information from social media and traditional media. The network created was a result of time-consuming extraction and processing of the different datasets. The network constructed includes nodes from news media and social networks like Facebook and Twitter.

First, the network was analysed using different graph metrics which helped better understand the diffusion processes and also gave an insight on where this kind of network stands compared to networks researched in other studies. Betweenness score indicated the high importance of articles in the network and similarly Facebook post but with a significantly lower score. Another metric considered was the vertex in- and out-degree which confirms the first indications of article and Facebook post importance. The degree distribution also showed that the network is scale-free which is common factor of social networks - on the contrary the assortativity score did not suggest network similarity to social networks.

Temporal patterns revealed many characteristics of this network - for one the fast pace of diffusion which tends to start immediately after the release of information and is active for 10-12 hours before the calming of the process with the highest peak occuring immediately after the release. Yang and Leskovec study [45] on the temporal patterns of news media phrases and Twitter hashtags showed similar results. Most of the information diffusion cascades had activity during the first 2 weeks after the initial release of source information. Temporal patterns implied some interesting surges as well - one of these was a small burst after 32 hours of the release where no exact reasoning could be found.

With topological patterns the common information flows of the network were analysed - for this different frequent subgraphs of various sizes were extracted. The patterns illustrate the central diffusion processes in the network and show that most of the diffusion processes are contained largely in one channel or domain. The central role of news articles and Facebook posts is once again confirmed by the topological patterns - these nodes in most cases are the sources of the diffusion cascades as well. Interesting motifs exhibited the diffusion patterns crossing the social media and traditional media domain borders - with the same type of nodes, articles and Facebook posts being at the periphery of domains and part of this kind of diffusion. The direction of the diffusion tends to be from news media to social media - the same phenomenon was found to be present in the study of event diffusion patterns by Kim et al. [21]. Large portion of the most common diffusion patterns are star-shaped which indicates the existence of multiple different routes for information to diffuse -

this kind of shapes were found to be dominant also in a study by Leskovec et al. [25] concentrating on one specific network, that is the blog network. An example distinction of the diffusion processes is given with a two level model - inter-domain diffusion is at the first level while the cross domain and channel diffusion is part of the second level processes.

This thesis demonstrates the extraction of diffusion patterns where the influence of different entities in the diffusion process is apparent. This confirms that the inclusion of multiple different bound networks in the study of information diffusion patterns help to understand the real world diffusion processes better and can be analysed in greater detail.

This thesis has provided an overview of information diffusion processes in a network comprised of multiple networks belonging to news media and social media domain. As well this thesis has provided an approach to study the diffusion processes from temporal and topological perspective. As an improvement, the approach taken here could be used for studying a network including more domains and domain channels. Additionally the impact of the patterns found could be analysed in more detail. Another future improvement plan could be perhaps the enhancement of the process of collecting the data, the network was built upon.

6 Acknowledgements

My Master's studies at the University of Tartu were supported by the the Estonian Information Technology Foundation and Skype Technologies OÜ.

References

[1] Python Programming Language. `http://www.python.org/`. [Online; accessed 10-February-2014].

[2] The R Project for Statistical Computing. `http://www.r-project.org/`. [Online; accessed 30-March-2014].

[3] yEd Graph Editor. `http://www.yworks.com/en/products_yed_about.html`. [Online; accessed 30-April-2014].

[4] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[5] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Sci. Am.*, 288(5):50–59, 2003.

[6] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674. ACM, 2011.

[7] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[8] Kainan Cui, Xiaolong Zheng, Daniel Dajun Zeng, Zhu Zhang, Chuan Luo, and Saike He. An empirical study of information diffusion in micro-blogging systems during emergency events. In *Web-Age Information Management*, pages 140–151. Springer, 2013.

[9] Fielding et al. Hypertext Transfer Protocol – HTTP/1.1. `http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html`. [Online; accessed 10-February-2014].

[10] Facebook. Facebook Comment. `https://developers.facebook.com/docs/plugins/comments`. [Online; accessed 29-January-2014].

[11] Facebook. Facebook Like. `https://developers.facebook.com/docs/plugins/like-button`. [Online; accessed 29-January-2014].

[12] Geoffrey Fowler. Facebook: One billion and counting. *The Wall Street Journal*, page B1, 2012. [Online; accessed 30-October-2013].

[13] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the Future*, 2012.

[14] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan Watts. The structural virality of online diffusion, 2013.

[15] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[16] Adrien Guille, Cécile Favre, Hakim Hacid, Djamel Abdelkader Zighed, et al. Sondy: An open source platform for social dynamics mining and analysis. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013.

[17] Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *SIGMOD Record*, 42(2):17, 2013.

[18] Marco A. Janssen. Simulating market dynamics: Interactions between consumer psychology and social networks. *Artificial Life*, 9:343–356, 2003.

[19] Nikhil S. Ketkar. Subdue: compression-based frequent pattern discovery in graph data. In *in OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 71–76. ACM Press, 2005.

[20] Riivo Kikas, Marlon Dumas, and Marton Karsai. Bursty egocentric network evolution in skype. *Social Netw. Analys. Mining*, pages 1393–1401, 2013.

[21] Minkyoung Kim, Lexing Xie, and Peter Christen. Event diffusion patterns in social media. In *ICWSM*, 2012.

[22] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[23] Joseph Kwon and Ingoo Han. Information diffusion with content crossover in online social media: An empirical analysis of the social transmission process in twitter. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 3292–3301. IEEE, 2013.

[24] Yong-Suk Kwon, Sang-Wook Kim, and Sunju Park. An analysis of information diffusion in the blog world. In *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, pages 27–30. ACM, 2009.

[25] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *Society of Applied and Industrial Mathematics: Data Mining (SDM07)*, 2007.

[26] Seung-Hwan Lim, Sang-Wook Kim, Soyoun Kim, and Sanghyun Park. Construction of a blog network based on information diffusion. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 937–941. ACM, 2011.

[27] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 6–14. ACM, 2012.

[28] Anastasia Mochalova and Alexandros Nanopoulos. On the role of centrality in information diffusion in social networks. In *ECIS*, page 101, 2013.

[29] Shahab Mokarizadeh, Peep Küngas, and Mihhail Matskin. Exploring information diffusion in network of semantically annotated web service interfaces. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, page 12. ACM, 2012.

[30] Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012.

[31] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701, Oct 2002.

[32] postimees.ee. Veerpalu moisteti oigeks. `http://sport.postimees.ee/1181708/veerpalu-moisteti-oigeks`. [Online; accessed 21-May-2014].

[33] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[34] Semiocast. Twitter reaches half a billion accounts more than 140 millions in the u.s. `http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US/`, 2012. [Online; accessed 27-October-2013].

[35] Dafna Shahaf, Jaewon Yang, Caroline Suen, Jeff Jacobs, Heidi Wang, and Jure Leskovec. Information cartography: creating zoomable, large-scale maps of information. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105. ACM, 2013.

[36] Xiaolin Shi, Belle L Tseng, and Lada A Adamic. Information diffusion in computer science citation networks. In *ICWSM*, 2009.

[37] Avaré Stewart, Ling Chen, Raluca Paiu, and Wolfgang Nejdl. Discovering information diffusion paths from blogosphere for online advertising. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, pages 46–54. ACM, 2007.

[38] Io Taxidou and Peter Fischer. Realtime analysis of information diffusion in social media. *Proceedings of the VLDB Endowment*, 6(12):1416–1421, 2013.

[39] Techopedia. Facebook Status. `http://www.techopedia.com/definition/15442/facebook-status`. [Online; accessed 29-January-2014].

[40] Twitter. Twitter retweet. `https://support.twitter.com/articles/20169873`. [Online; accessed 10-February-2014].

[41] Twitter. Twitter Tweet. `https://support.twitter.com/articles/166337-the-twitter-glossary`. [Online; accessed 29-January-2014].

[42] Marina von Steinkirch. Information diffusion in twitter. 2012.

[43] Bo Xu and Lu Liu. Information diffusion through online social networks. In *Emergency Management and Management Sciences (ICEMMS), 2010 IEEE International Conference on*, pages 53–56. IEEE, 2010.

[44] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 599–608. IEEE, 2010.

[45] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.

[46] Jiang Yang and Scott Counts. Comparing information diffusion structure in weblogs and microblogs. In *ICWSM*, 2010.

A Appendices

A.1 Appendix 1

Python script for mining URLs in news article and blog post dumps.

A.2 Appendix 2

Python script for extracting Facebook different entity relations and mining URLs in Facebook posts and comments

A.3 Appendix 3

Python script to extract Twitter tweet relations and URLs present in the tweets.

A.4 Appendix 4

Python script to extract article comment relations and URLs present in the article comments.

A.5 Appendix 5

A script written using Python to make HTTP requests to validate URLs.

A.6 Appendix 6

A script for constructing the graph in edge list format.

A.7 Appendix 7

Includes different R scripts to analyse the network.

A.8 Appendix 8

Different scripts to carry out frequent subgraph mining using SUBDUE.

A.9 Appendix 9

Python scripts to calculate some of the network metrics.

A.10 License

Non-exclusive licence to reproduce thesis and make thesis public

I, Oliver Soop
(date of birth: 03. December 1988),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Local Information Diffusion Patterns in Social and Traditional Media: The Estonian Case Study

supervised by Ph.D. Peep Küngas

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **26.05.2014**