Chances and challenges of computational data gathering and analysis: The case of issue-attention cycles on Facebook

Niina Niskala
PhD Candidate; Department of Communication, University of Jyväskylä; Ullantorppa 10B29, FI-02750 Espoo, Finland; +358407249225; niina.niskala@jyu.fi (corresponding author)

Jukka Rohila
M.Sc. (Econ. & B.A.); Senior Consultant, Siili Solutions Oyj; Arkadiankatu 4–6, FI-00100 Helsinki, Finland; +358408206767; jukka.rohila@siili.fi

Epp Lauk
Professor, PhD; Department of Communication, University of Jyväskylä; Mesikämmen 6A2, FI-40400 Jyväskylä, Finland; +3585767860; epp.lauk@jyu.fi

Turo Uskali
Senior Research Scholar, PhD, Docent; Department of Communication; University of Jyväskylä; Keskussairaalantie 4 (Building L), FI-40014 Unversity of Jyväskylä, Finland; +358 405488448; turo.i.uskali@jyu.fi

Jukka Jouhki
Senior Lecturer, PhD, Docent; Department of History and Ethnology; University of Jyväskylä; P.O. Box 35 (Building H), FI-40014 University of Jyväskylä. Finland; +358505750576; jukka.jouhki@jyu.fi

Maija Penttinen
Undergraduate Student; Department of History and Ethnology; University of Jyväskylä; Voionmaankatu 9 C 58, FI-40700 Jyväskylä, Finland; +358504110664; maija.s.penttinen@student.jyu.fi

Research conducted at University of Jyväskylä.

This work was in part supported by the Media Industry Research Foundation of Finland and Faculty of Humanities, University of Jyväskylä.

Words count: Core text (Title–References) 7677, Tables and Figure captions 1099.

Digital and social media and large available data sets generate various new possibilities and challenges for doing research focused on perpetually developing online news ecosystems. This paper presents a novel computational technique for gathering and processing large quantities of data from Facebook. We demonstrate how to use this technique for detecting and analyzing issue-attention cycles and news flows in Facebook groups and pages. Although the paper concentrates on a Finnish Facebook group as a case study, the method demonstrated can be used for gathering and analyzing large sets of data from various social network sites and national contexts. The paper also discusses Facebook platform regulations of data gathering and ethical issues of doing online research.

Keywords: Facebook; Computational data gathering; data warehouse; digital and social media research; hybrid news ecosystem; issue-attention cycle; news flows; semi-public data

Introduction

Alongside the digital revolution and development of virtual environments, social scientific research is experiencing a paradigm shift towards computational approaches (Chang, Kauffman, and Kwon 2013, 67). Traditional qualitative and quantitative methods of social sciences appear to be limited in studying the phenomena of such rapidly altering environments as the internet or social media (Karpf 2012, 646). As social science research begins to use data-driven methods and novel tools for data gathering and analysis, the multidisciplinary approach is increasingly common (Chang, Kauffman, and Kwon 2013, 70). According to Steensen and Ahva (2014), we are living through the 'fourth wave' of research on digital journalism, which following after normative, empirical and constructivist waves emerged mainly because of the new practices related to the social media. While arguing there is a need to reassess the theories, they do not, surprisingly, identify any similar need for the new methodologies.

Internet and social media entail various platforms and social network sites (SNS), which have not yet been either sufficiently harnessed for research or have had their potential discussed, especially from the perspective of data gathering and analysis for social scientific or specifically journalism research. For example, Facebook generates a hybrid digital communication and news ecosystem where issues rise and fall in newsfeeds and on specific groups and pages, and news flows are continuously created by the users themselves, including sharing news generated by online news media.

Consequently, internet and social media enable the generation and access of massive data sets ('big data') as sources for research material (Bollier 2010). Although there is a lack of common agreement and clarity of the definition of 'big data' (cf. Boyd and Crawford 2012; Ukkonen 2013), the term has been widely adapted by media and journalism scholars (Couldry and Powell 2014; Lewis 2014; Lewis and Westlund 2014). Large digital data sets used in data journalism are also being referred to as 'open data' "that can be freely used, re-used and redistributed by anyone" (Open Knowledge Foundation 2012; Mair et al. 2013; Coddington 2014; Parasie 2014). Conventional journalism research benefits from using 'big data' as replacement to traditional representative sampling (Couldry and Powell 2014, 1). The current trend towards ubiquitous and mobile communications naturally increases the variety of 'big data'. The challenge is how to channel these data streams into knowledge, journalism or research (Lewis and Westlund 2014, 4).

The objective of this paper is to demonstrate the building procedure and possibilities of a new computational approach for gathering and processing online 'big'data' and show how it can

be used to detect issue-attention cycles and news flows in Facebook groups and pages. We use a case example from Finnish Facebook Community page "Valio out of Fennovoima's nuclear plant project" (2014). This is a group protesting against cooperation between the food producing company Valio and the nuclear power company Fennovoima in building a new nuclear power plant in Finland. Facebook has, in Finland, become the main channel of the public's online activities and news consumption. A substantial majority of the population – about 86 percent – regularly uses the internet, and Facebook is used by 95 percent of all Finnish SNS users (Statistics Finland 2014). The case, as a topical issue in Finnish public debate, is particularly interesting from the journalism study perspective and will be used for detecting how the data gathered from Facebook can be used to compare and find synergy between public's attention waves and media influence and operations.

Online news and 'issue-attention cycles'

Online news is characterized as a revolutionary hybrid news medium (e.g. Allan 2006; Kautsky and Widholm 2008), which affects the role of journalists as mediators and moderators of information (Hermida et al. 2012) as well as changes the audience's behaviour. Social media platforms, such as Facebook and Twitter, form a digital news hybrid with traditional news media services. Facebook is acknowledged as one of the primary vehicles for news flows and exposure to news (Baresch et al. 2011; Bell 2014). News, via Facebook, is nowadays a mixture of Facebook users' posts, Facebook groups' and (fan) pages' posts, as well as advertisers' messages.

Consequently, in contrast to passive audiences, people can now also be considered as mediators and gatekeepers of news online. However, the audience's attention cannot be taken for granted. "Instead of a traditional push-model, users are free to navigate between sites to seek the information they desire and select their own versions of the daily news" (Weber and Monge 2011, 1063). Thus, in contrast to the traditional procedure of newspapers gathering their audiences, Baresch et al. (2011, 2) refer to "a new kind of news consumption strategy, a new kind of consumer, a 'stumbler', so to speak, who gets nearly all his or her news through incidental or socially selected exposure". Online news sites and portals fiercely compete to be the quickest at catching the audience's attention, thus publishing breaking news almost in real time. The same news criteria as in traditional media largely become irrelevant, since the virtual space is practically unlimited. Social media largely serves as a distributor of the news, but also as an independent news source. The dissemination of issues is unpredictable and uncontrolled, as everyone can at any time 'share' a link, or news, to any number of SNS (social network sites) and other platforms. In addition, algorithms produce and distribute news online that can also be redistributed by social media users.

Moreover, the attention span of the media and the public on an issue is not unlimited. The cyclic character of public attention was noticed already in the pre-digital media era. More than 40 years ago, Anthony Downs outlined a concept of 'issue-attention cycles' characterizing the rising and fading of public attention and concern towards major societal issues (Downs 1972).

Downs (1972, 39-40) suggests five stages of 'issue-attention cycles', characteristic to the American society and media of the time: 1) the pre-problem stage, where only some experts or interest groups are aware of the problem; 2) alarmed discovery and euphoric enthusiasm, where the public "becomes both aware of and alarmed about the evils of a particular problem; 3) realizing the cost of significant progress, which includes major sacrifices by large groups of

population; 4) gradual decline of intense public interest and enthusiasm; 5) the post-problem stage, where the issue moves into a "twilight realm of lesser attention or spasmodic recurrences of interest". In the digital environment, issues rise and decline rapidly, but the attention cycles are not necessarily shorter than in traditional media as Anderson et al. (2012) demonstrate. In the pre-internet time, the traditional media, in fact, governed the issue-attention cycle. The structure of the news flows in the traditional media is based on newsworthiness and the space the topical stories get in print and broadcasting outlets. However, before the era of computational approaches and data-driven methods, detecting and explaining the cyclic nature of public attention and major issues, which appear in the combined news flows of the digital news ecosystem, was not feasible.

Approaches to computational data gathering for research

Using semi-public data (data collected with a user account) to retrieve data for Facebook group and page content analysis research with a computational approach such as the one presented in this paper, i.e. a tool using Facebook's own Application Protocol Interfaces (API) to gather all available communication activity data from the platform's pages and groups and organizing it into a warehouse, is still very rare in social sciences. Semenov (2013) discusses many aspects of social media data analysis, implementation and repository designed for monitoring communities on social media sites (see also Semenov, Veijalainen, and Boukhanovsky 2011). In their working paper Zlatanov and Koleva (2014) are using a software application called Opinion Crawler designed to extract data from open Facebook groups and use it for data analysis through people centric models and text network analysis connected to online originated protests. Nevertheless, the objectives of the aforementioned studies are quite different from this paper's; they do not explain the software design specifics nor data organization procedures of Facebook data per se and are not done from the journalism study perspective.

Indeed, many Facebook data collection applications and technical platforms are available online for researchers (e.g. Digital Footprints, NodeXL, Netvizz, RFacebook, SocialMediaMineR, and Facepager). For example, Netvizz (2015) is a readymade application tool for retrieving data from Facebook, which asks for target users' permission to access their public profile and friend lists. Though quite similar in providing lists of posts and their likes and comments, in contrast to our approach it relies on the researcher to be a member of a group or liking a page, and is dependent on the tool creator's decisions on output data. The tool provides less data and analysis possibilities compared to independent data retrieval and building one's own data warehouse.

Another more technical and nearly identical data tool to ours in its data retrieval method is Facepager (2015), which fetches publicly available data from Facebook, Twitter and other platforms with an open standard format for transmitting data (JSON-based API) and stores it in a database. This tool can gather all the Facebook data, but in an unorganized form. Our approach is to transfer the data into the data warehouse in a specific model schema, which helps to organize and analyze the data.

Facebook and Data Collection

Privacy settings greatly impact on the results of data gathering on Facebook, excluding information the user has decided to hide from others (Giglietto, Rossi, and Bennato 2012). For

example, analyzing newsfeeds on Facebook could turn out to be more difficult than originally expected due to the general notion of only sharing the timeline with a limited list of 'friends'. Instead, focusing the research on Facebook pages could improve the results of data collection, because Facebook pages are regarded as public material with no limitations to its internal data (Giglietto, Rossi, and Bennato 2012, 152).

While conducting research on Facebook, it is important not to violate the common principles of the site. The general notion in Facebook's 'Principles' and 'Statement of Rights and Responsibilities (SRR)' about the data available focuses on the possibility of the users to individually determine the information they are willing to share publicly. If users do not limit the availability of information, it becomes public data, and Facebook is not responsible for what information received from the site is used (see Facebook 2014a; 2014d.)

Generally, access to SNS can be categorized into three data: *public data, semi-public data* and *dark data*. Public data can be retrieved from public interfaces without signing a user agreement with a service provider. Semi-public data can be accessed from public interfaces by signing a user agreement and using the user account to retrieve data. Dark data is obtained with techniques that the service provider has not intended for use or are against the user agreement of the service. For example Facebook provides more data in their web user interface than from their APIs. This difference has given rise to software tools like 'scrapers' that instead of using machine interfaces use interfaces meant for human users to pool more detailed information from the network. Other 'dark' tools, like harvesters, robots and spiders, gather data blindly in an attempt to remodel the social network. Facebook's RSS and 'Automated Data Collection Terms' (ADCT) dictate terms of using automated tools, such as harvesting bots, robots, spiders, or scrapers, and indicate a need to ask Facebook's written permission for using such tools or storing data, and forbid using any acquired data for business or advertisement purposes (Facebook 2014d; 2014h). Operators of Facebook's own Platform applications or websites, and users of Social Plugins must comply with the 'Facebook Platform Policy' (FPP) (Facebook 2014g)

Below we discuss the collection of semi-public data from users and communication data from Facebook open groups and pages, and use the same data from a page in the case study analyses. During the creation of the data gathering techniques, special emphasis was put on acting in accordance with Facebook's user agreements, principles, and any intent implied in the SRR, FPP, and ADCT. Particular attention was paid to developer rules, protection of data, IDs and not selling or reproducing the data, applicable to our computational approach of semi-public data gathering, as we are not developing a specific platform app or website nor using the indicated automated data collection tools.

Building Data-Centric Research Approaches for Studying Facebook

We describe here the path that was taken to create a research tool for studying Facebook. Facebook currently offers a readily available user interface data tool and a simple content search system, the Graph API Explorer. It is a software environment created for third-party developers, where it is possible to create applications to access data on Facebook by asking permission from users (Facebook 2014c). Instead of using the Graph API Explorer, we gathered semi-public data from pages and groups by using Facebook's own APIs and public interfaces, which allowed us broader freedom of research and more data to be obtained, still following the general (developer) principles and regulations of Facebook.

We will initially define key concepts and motivations behind using computational data gathering, then explore the possibilities of data organization, warehousing and analysis, and in the following section, demonstrate some basic case examples of the application of research data in studying issue-attention cycles and news flows in Facebook pages. While describing the building process of the tools, we have focused more on concepts and problem solving than on a direct hands-on-approach. Our understanding is that conducting data-centric research and building computational tools is not purely technical work – it is more about thinking on what data we can obtain and what its meaning is from the point of research.

Automatic vs. Manual Data Gathering and Some Basic Concepts

The biggest drawback of manual data gathering is that this method is slow and prone to human errors. The Facebook page used in this article as a case example had 966 posts to its feed and 1,921 comments attached to these posts. In addition, manual method makes only a limited amount of all data available. While Facebook shows the key information in a web or mobile client, there is more data available from the system's APIs, such as metadata on how the system handled the content in question. Furthermore, during the lengthy manual data gathering, the information available can change or it can become unavailable. Although computational data gathering does not completely remove the risk of data being changed or going missing, it minimizes the window of opportunity where it can happen.

For an avid user, who has used Facebook public web or mobile client, the concepts, terms and scopes might be self-evident, but not necessarily from their technical point of view. Depending on what API is in use, there are various terms for the underlying data and restrictions that are invisible for users when browsing the service. Some of the main concepts are explained in Table 1.

TABLE 1. NEAR HERE

Searching and Retrieving Data from Public Interfaces

Both public interfaces of Facebook, the Graph API and FQL, enable keyword search of selected data in the system. Currently in the version 2.2 it is possible by using the Graph API to make a keyword search according to names of users, pages, events, groups, places and locations (Facebook 2014e). Previously it was also possible to search posts with keyword search but unfortunately this feature is being removed from the system with the introduction of version 2.2 and the deprecation of the 2.0 version of the platform. With the same upgrade, Facebook is also removing support for FQL (Facebook 2014b).

After a search is made (for the how, see Facebook 2014e), the system returns a list of matching results. Only users who have not removed themselves from public search are reached. When a researcher finds an interesting source of data and this data is publicly available, the data is retrieved by issuing calls to different endpoints of the Graph API.

There are, however, a few restrictions on which data can be retrieved. First of all, if the researcher is not the owner or administrator of a feed, no data about who has liked the page or group is available. Secondly, the biggest restriction is the API call limit. Currently Facebook only allows 600 calls per 600 seconds per authentication token, i.e. in simple terms one call per second can be issued to Facebook (see Mangobug 2012). While this restriction, for a human user,

would not be a limiting one, for a computational data search and retrieval system this is a major hindrance and limitation. Also, while it takes only around 40 seconds to retrieve 1000 posts from a feed, it will take much more time to get other data, such as comments on the posts. In other words, while access to basic items is fast, being able to gather all data takes much longer. Thus, when designing an application to search and retrieve data, researchers should make an effort to classify the kinds of data they want before embarking on building a data storage and warehouse tool.

Push-Stream and Pull-Stream Views on Data Retrieve

Facebook's owners have access not only to what people do on their pages, but they also have information about who has viewed what and for how long. Such an overall *datascape* could be called *whirlpool*, where events external to the social network create internal actions in the system, which then can again create actions outside the network.

The 'whirlpool' can be further broken down to two different ways of looking at retrievable data from Facebook. When a user makes a post, comment, like or share, they actually generate a messaging event to initiate different actions in the system. This view of retrievable data could be called the *push-stream*: user generated events push the system to perform different actions. Another view of Facebook would be constructed on the basis of what it messages to a user. Individual users have their own event wall or feed that is filled by Facebook from the content generated by a particular users' network of friends and by advertisements displayed by the system itself. This view of the retrievable data could be described best as the *pull-stream* where individual users pull content from the system by browsing the content of their own or other users' walls.

However, Facebook sets restrictions on what is possible to retrieve from the system. Logged users can only access their own feeds, feeds of their own network, feeds of the other users who have set their feeds to public and feeds of open groups and pages. In addition, Facebook does not give any information about individual visits to the content of a feed, besides for administrators of pages. These restrictions alone make it impossible to retrieve pull-stream data. In addition, Facebook does not give detailed information about 'shares' and 'likes'. In the case of 'shares' information about who shared and to whom is missing. The only statistic the system gives out is the total number of 'shares'. In the case of 'likes', the system does not give the time of an individual 'like'. These restrictions hinder retrieving push-stream data, but do not make it completely impossible.

In conclusion, we can see that from the data that Facebook provides to third parties, the push-stream forms the most complete retrievable data. The pull-stream based view is currently impossible to be reconstructed and the aggregated level is still heavily restricted. However, a limited 'whirlpool' view could be constructed by combining push-stream data with aggregated pull-stream data as well as with external sources, such as traditional media. However, the recommendation based on our experience is to concentrate on retrieving data from the push-stream, looking at actions and their actionable effects.

Need for Clarifying Data and Creating a Data Warehouse

When testing the computational data retrieval approach, we noticed multiple practical problems. Firstly, social networks as other IT-systems are under constant change where new

features are developed and old techniques are terminated. Also, alongside changes in the official APIs, other changes e.g. the privacy policy could lead to alterations of available data. Thus, the research subject is a moving target that can suddenly change, in the worst case denying access or changing key functionalities and affecting the research results.

Another problem is that the APIs are intended to extend the functionality of their respected services. They are not designed for data analysis but for the day-to-day system operations. For example, each 'post' retrieved via Graph API has 28 attributes that describe it and its relationship to other items and system metadata. Also, there are overall 31 endpoints in the system from where different data can be retrieved. Thus, there are too many different data objects and attributes to handle without any categorization or connection clues.

These problems evoked the idea that the data retrieved should be logically separated and isolated from the data format provided by the source system, and when moved into a different system for analysis, the data should be coded into simpler form. Thus, a separate data warehouse was coded and constructed to manage the data.

Data warehouse in computing refers to a system that is used for saving data from one or multiple source systems into a single set for retrieving data reports and analysis. There are two different design philosophies regarding data warehousing. The first is the dimensional approach of Ralph Kimball's star schema where measurable and quantitative data are stored in fact tables, whereas descriptive attributes related to the fact data are stored in dimensional tables (Kimball 1996). The second design approach is Bill Inmon's 3NF model (Third Normal Form) where data is structured as much as possible in order to minimize data redundancy through normalization (Inmon 1992). It is also possible to make a combined approach of the two models. Hence, we decided to combine the idea of the star schema (storing the data into separate tables) with data normalization (simplifying and combining similar forms of data into the tables), thus forming a data model that is easy to search and analyze.

The data warehouse of this study was built on the basis of a ***push-stream based event data centric model*** modelled into a star schema, but structured and normalized as much as possible. 'Event' was defined to be a time-associated transaction in the system. Due to Facebook's restrictions on accessing data, only 'posts' and 'comments' could be defined as 'events', and 'likes' and aggregated data about 'shares' was used to describe 'events' to which they were tied. The data model used was based on a dimensional approach where 'posts' and 'comments' would be stored as 'events' and all the other information describing and relating to them stored to dimensional tables (see Table 2 and Figure 1).

TABLE 2. NEAR HERE

FIGURE 1. NEAR HERE

The technical reason not to store 'like' data to an 'event' table was to make it smaller and simpler. Especially in feeds of groups and pages, every 'post' and 'comment' has multiple 'likes' and thus, the size of the event table would grow immensely. Queries would also become more complex as 'likes' would be treated as child events of 'posts' and 'comments' versus a direct relation linkage via the 'like' table.

In addition to simplifying the form of the data, clarification and warehousing also enable using a single set of analytical tools to address multiple social networks via a common data warehouse. Although data from different social networks differ, their functionality in the

conceptual level is similar (see Table 3). By identifying similarities between social networks, data from multiple systems can be brought into a single system and transcoded into a single format, enabling researchers to build and use exactly the same analytical tools for both networks.

TABLE 3. NEAR HERE

Possibilities of Analyzing Data from the Warehouse

After the decisions and organization of data, the warehouse now stores the data in five tables (see Figure 1): f_events, d_content, d_entity, d_event_type and r_likes. From these tables both the d_content and d_event_type tables describe what was the content of an event, thus together forming content centric type of the data. Tables r_like and d_entity deal with actors and their relation to events and each other, forming people centric type of the data. The last type of data is the 'event' based, consisting of 'posts' and comments made in a feed.

The content centric data is the most unrestricted from the three types. Content in general can be retrieved without hindrance. Analysis of the content of texts, photos and videos is still the major domain of qualitative research. Content can be, indeed, automatically analysed for example, with sentiment analysis, using word lists to calculate the content of a text, but these techniques are yet under construction and their usefulness is under discussion. However, the most useful function of storing the content along with other feed information is the ability to get the content out quickly and with additional information, such as time and date, number of likes, comments, users tied to a specific content, and so forth.

The people centric data is a rather limited type, not only due to Facebook's restrictions to personal data, but because the user information available is tied to user generated events. If a person has read a post or a comment but not done any other activity, there is no trace of that user. Thus, the user information is limited only to people who at least once were active in a feed. The information always available about the users includes first name and last name, used language version and user's profile picture. Gender information is available in 99.9 percent of cases, friend count in 54.7 percent, and affiliation information in 2.9 percent of cases. This is essentially all the information that Facebook gives about users directly, but it does enable making comparisons based on gender and on amount of friends a user has.

The event centric data enables more data to be gathered and analyzed. In a context of a feed, one can track a user's total number of 'posts', 'comments', 'likes', and 'shares' and commented, liked and shared posts and comments. Further information can be generated by taking into account the type of content and time of a post or comment in question. The problem with the event centric data, as noted before, is that Facebook does not give information about individual shares and about the time when a 'like' was made. Thus, analysis of the event data is analysis of actions and their responses. For example, a 'like' is always a response, while a 'comment' and a 'post' can be both actions and responses to other posts and comments. With this data, we can generate a view of what has happened, what are the responses to it, and by combining this data with data about content and users, we can start to explain behavior of a feed.

Using the Computational Approach in Journalistic Research: Case Facebook

To demonstrate potentials of the Facebook data and its organization for journalistic research, we chose a page of a protest group called "*Valio out of Fennovoima's nuclear plant*

project" (2014) with 3,095 followers. The page was searched from Facebook with Graph API, all its available semi-public data was retrieved and directed to the organized data warehouse, which automatically saved the data to the assigned tables.

In the following, we use the concept of the 'issue-attention cycle' for demonstrating the potentials and possibilities of the computational data gathering. Our aim is not to attempt to exhaustively explain the ebb and flow of public attention on the issue of building a nuclear power station in Finland and all the activities of the respective Facebook group. Instead, we are trying to show how to discover and visualize the waves of attention using large data sets, and indicate some possible basic ways of analyzing them.

Detecting Issue-Attention Cycles

Groups and pages formed to support or protest against an issue make it possible for journalism and media researchers to observe issue-attention cycles of certain societally topical issues. Our example consists of above mentioned group's data since its birth, from week 32 of 2011 to week 45 of 2014.

Focusing on the page's event centric data of 'posts' and related communication activity, such as shares, likes and comments, it is possible to observe the intensity of the group's attention. Table 4 shows a data table example that contains weekly downloads of the page's 'posts' and post related activities. It is also possible to take a daily download of 'posts' for even more specific cycle evaluations.

Table 4 firstly shows the *year, month* and *week* of the 'posts', then the total amount of *post*s grouped by time, and also by their producer: *post by source* indicating the page/administrator of the page as the source, *post by other* indicating any other actor as the source, and total number of *posters*, i.e. amount of individual producers of the posts. The Table also shows the same total amounts and grouped information of post related *shares, likes, comments* and *comment likes*.

TABLE 4. NEAR HERE

By categorizing the activity information according to each producer, one may find some interesting aspects. For example, in this short data excerpt we can see that in total 103 'posts' were made on the page wall, and most of them by the page/administrator, i.e. source (n=89, 86%). In addition, there are no shares made by the group 'members' (i.e. other), only by the page and with large quantities, and nearly the same applies to 'likes'. This may give an initial indication of the group's internal dynamics and objectives.

From the issue-attention cycle perspective, the total amounts of posts, shares, likes, and comments (data columns highlighted in Table 4) give a good overall picture of the communication activities of a Facebook group. Figure 2 visualizes the overall data set with a specific focus on total amounts of 'posts' and their tied activities of shares, likes, and comments, showing a synopsis of the amounts from approximately every other week from years 2011-2014.

FIGURE 2. NEAR HERE

By looking at Figure 2, we can make observations of the issue-attention cycle in the social media context. The total data figure shows how attention has been relatively steady and

low-level immediately after the launch of the group (during 2011-2012) and only towards the end of 2013 and in 2014 have activities been boosted. The early phases of the formation of the group may reflect the 'pre-problem' stage, as the initiators of the group must have been people who had enough information to be worried about the issue. Alongside the growing attention of the traditional media (close to the end of 2013), when *Fennovoima* made a deal with the Russian nuclear energy corporation *Rosatom* (Taloussanomat, Dec. 21, 2013), the group's activities reach the 'alarmed discovery' stage. The activity increases even further in the beginning of 2014. At this time, media attention focused on the fact that the Fennovoima's plant project had, due to funding problems, been transformed into a Russian project (e.g. YLE, March 27, 2014). During the 'alarmed discovery' stage, the group actively shares information, which also gets significantly large quantities of 'shares' and 'likes'.

To be able to give an exhaustive explanation of the actual reasons of the activity peaks, their relation with the public agenda and the group's inner development, qualitative analysis is also necessary. The most lucrative way of starting qualitative analysis is by uploading content centric data of posts from the warehouse for content analysis. This would, for example, reveal the specific content of the posts and comments that created the high activity peaks of weeks 12-14, 2014. The quantitative data enabling activity visualization is, nevertheless, a good foundation for any further analysis.

News Flows as Escalators of Attention

Communication of groups and pages on Facebook includes links to online news articles and contents of other social media and websites, as well as other users' comments, shares and 'likes' on the links, which all are components of social media's news flows. Examining these news flows by tracking, for example, articles on particular topics, changing patterns of the consumption and production of news can be described, as well as issue attention cycles explained.

Focusing on content centric data of the links attached to the 'posts' shows what news articles or other content, and from which sources have been linked. In addition, the event centric data of related comments, likes and shares shows how the news links generate activity and thus escalate their impact and create new news flows.

Table 5 shows a snapshot of one way of retrieving and organizing the data table that is formable by this focus from the warehouse. The complete data table includes 121 links to sources retrieved from the group's wall since the time the group was formed, from week 32 of 2011 to week 45 of 2014.

More specifically, by looking at the first *source* column row 'www.facebook.com', the second column *referred* indicates the total amount that a link from the source 'www.facebook.com' has been posted on the page (n=222). The third column *referrers* indicates the total number of actors posting the link (n=9), and the next three columns categorize the *referred* information according to the specific producing actor: *users*, i.e. individual people (n=6), Facebook *page* (n=3) and Facebook *group* (n=0). The data Table's six final columns show the *like count*, i.e. total number of likes made on the 'Facebook' source links (n=3661), their total *share count* (n=1279) and total *comment count* (n=1038), and the average frequencies each link has been *liked (avg)*, *shared (avg)* and *commented on (avg)*.

TABLE 5. NEAR HERE

The data output includes a lot of specific producer information, which offers interesting perspectives for interpretation, but also gives the basic total amounts of used' links' and their likes, shares and comment counts. The overall data offers vast possibilities of counting and correlations. For example, one simple option is to start by counting news source links in comparison to other used link sources (e.g. entertainment) to evaluate media influence, proceeding to the comparison of different media sources. Our study's results show that the most linked source on the page has been Facebook itself (n=222), the second most used source has been YLE (Finnish National Public Service Broadcasting Company) (n=57) and the third was Kaleva (a daily newspaper) (n=40). Interestingly, Finland's most popular and authoritative quality newspaper *Helsingin Sanomat* appears only as the fifth source.

In addition, by focusing on the total amounts of attention and activity the posted links have generated, i.e. likes, shares and comments, one can measure the general impact of the links. Figure 3 shows the four most linked news sources and their escalated activity.

FIGURE 3. NEAR HERE

One could continue by making more complex analyses and models of the way the amounts and content of posted news links affect the way they have gained attention among the public (likes, comments) and get to be forwarded to new publics (shares).

The described approach offers various possibilities for setting the research focus and analyzing the data. Quantitative analysis can be combined with qualitative analysis, for example, by retrieving content centric data with a focus on time and full content of the links, as the links can be opened and qualitatively analyzed. When looking at the cycles of issue attention, one might for example, choose links from the high attention period, study their journalistic framing and narratives, and compare them to the low attention periods.

Conclusions and Discussion

This paper is a reflection to the paradigm shift in social scientific research towards a computational approach in data gathering and analysis. Our main conclusion is that use of innovative research techniques of the internet studies and analysis of large data sets in studying the digital and social media enormously widens the scope and quality of the information that social scientists can have in their disposal. The traditional methods of social sciences, such as surveys, various ways of text analysis, interviews etc., are not sufficient for studying new kinds of information streams of the new hybrid digital news ecosystem that combines the online media, social media and other digital sources of information/news. They remain limited also in researching the consumers and producers acting in this ecosystem, as they, too, become increasingly combined (often called as 'prosumers'). The altering subjects and focuses of research also require enlargement of variety of information available for researchers.

In this article we introduced a new option for gathering and processing large data sets for studying attention and information flows on SNS, specifically Facebook. Much of the data Facebook contains are not freely accessible, but there are healthy amounts of open data for example on Facebook pages and open groups. One way of computationally accessing such data is to search and gather semi-public data by using Facebook's own APIs and public interfaces. This method allows more data to be obtained and freedom of data organization compared to

specific applications asking permission or other readymade online tools. In the approach we emphasized isolating the data from the source system and coding it to a simpler form in a separate data warehouse. Before building a data warehouse model, it is nevertheless important to understand the aspects of the available data and their connections, which are presented in this paper. Understanding the technical aspects is also helpful in retrieving and planning data analysis (beyond the case examples of this paper).

The case example analysis in this paper was kept quite simple and limited to the 'issue-attention cycle' of one topical issue in the Finnish public debate and link flows on the group page. The analysis showed how large data sets can be used to present time scales of high and low waves of online activity and attention on an issue. By comparing them to real societal happenings and established media content analysis, various cycles and explanation of issue-attention are possible to outline. In addition, the process allows for tracking news flows on social media and their impact among the public both by quantitatively comparing how much shared news get attention online and qualitatively analyzing for example, comments on shared news articles. What news and how they gain public's attention and become forwarded in social media provide journalists and media houses with valuable information regarding online news consumption and news flows.

The same data gathering and processing technique can be used for studying various other journalistic aspects both quantitatively and qualitatively, including integration of other SNS. For example, comparing traditional media content analysis with online public's societal or political issue attention and framing may give new insight in who sets the agendas in todays' society; online publics or the media? In addition, the data can be used to detect how media use online topics as sources compared to online public finding their topics from traditional/online media.

The ethical rules and views of collecting and using online data for research still remain under debate. However, by abiding by general laws and the rules of the platform or SNS under scrutiny, using case-by-case reflection, and securing the anonymity and safety of individuals and data, a researcher should be able to conduct ethically acceptable online research.

References

Allan, Stuart. 2006. *Online News: Journalism and the Internet*. Maidenhead: Open University Press.

Anderson, Ashley A., Dominique Brossard, and Dietram A. Scheufele. 2012. "News Coverage of Controversial Emerging Technologies: Evidence for the Issue Attention Cycle in Print and Online Media." *Politics and the Life Sciences* 31 (1–2): 87–97.

Baresch, Brian, Lewis Knight, Dustin Harp, and Carolyn Yaschur. 2011. "Friends Who Choose Your News: An Analysis of Content Links on Facebook." Paper presented at the International Symposium on Online Journalism, Austin, Texas, April 1–2. Accessed November 15, 2014. https://online.journalism.utexas.edu/2011/papers/Baresch2011.pdf.

Bell, Emily. 2014. "Silicon Valley and Journalism: Make up or Break up?" The Reuters Memorial Lecture 2014 for the Reuters Institute in Oxford University, November 21st. Accessed

November 24, 2014. https://reutersinstitute.politics.ox.ac.uk/news/silicon-valley-and-journalism-make-or-break.

Bollier, David. 2010. *The Promise and Peril of Big Data*. Washington, DC: The Aspen Institute. http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.

Boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication, & Society* 15 (5): 662–679.

Chang, Ray M., Robert J. Kauffman, and YoungOk Kwon. 2014. "Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data." *Decision Support Systems* 63: 67–80. doi:10.1016/j.dss.2013.08.008.

Coddington, Mark. 2014. "Clarifying Journalism's Quantitative Turn: A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-Assisted Reporting." *Digital Journalism*. doi:10.1080/21670811.2014.976400.

Couldry, Nick and Allison Powell. 2014. "Big Data from the Bottom Up." *Big Data & Society* 1 (2): 1–5. doi: 10.1177/2053951714539277.

Downs, Anthony. 1972. "Up and Down with Ecology: The Issue Attention Cycle." *The Public Interest* 28: 38–51.

Facebook. 2014a. "Principles." Accessed November 7, 2014. https://www.facebook.com/principles.php.

Facebook. 2014b. "Platform Upgrade Guide." Accessed December 18, 2014. https://developers.facebook.com/docs/apps/upgrading/.

Facebook. 2014c. "Graph API Explorer." Accessed November 21, 2014. https://developers.facebook.com/tools/explorer/145634995501895/.

Facebook. 2014d. "Statement of Rights and Responsibilities." Accessed November 7, 2014. https://www.facebook.com/legal/terms.

Facebook. 2014e. "Using the Graph API." Accessed November 21, 2014. https://developers.facebook.com/docs/graph-api/using-graph-api/v2.2#fieldexpansion.

Facebook. 2014f. "Facebook Query Language (FQL) Overview". Accessed March 16, 2015. https://developers.facebook.com/docs/technical-guides/fql.

Facebook. 2014g. "Facebook Platform Policy". Accessed March 16, 2015. https://developers.facebook.com/policy/

Facebook. 2014h. “Automated Data Collection Terms”. Accessed March 20, 2015. https://www.facebook.com/apps/site_scraping_tos_terms.php.

Facepager. 2015. Accessed March 18, 2015. https://github.com/strohne/Facepager.

Giglietto, Fabio, Luca Rossi, and Davide Bennato. 2012. “The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source.” *Journal of Technology in Human Services* 30 (3–4): 145-159. doi:10.1080/15228835.2012.743797.

Hermida, Alfred, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. “Share, Like, Recommend: Decoding the Social Media News Consumer.” *Journalism Studies* 13 (5–6): 815–824. doi:10.1080/1461670X.2012.664430.

Inmon, William H.1992. *Building the Data Warehouse*. NY: John Wiley & Sons, Inc.

Karpf, David. 2012. “Social Science Research Methods in Internet Time.” *Information, Communication & Society* 15 (5): 639–661. doi:10.1080/1369118X.2012.665468.

Kautsky, Robert and Andreas Widholm. 2008. “Online Methodology: Analysing News Flows of Online Journalism.” *Westminster Papers in Communication and Culture* 5 (2): 81–97.

Kimball, Ralph. 1996. *The Data Warehouse Toolkit. Practical Techniques for Building Dimensional Data Warehouses*. NY: John Wiley & Sons.

Lewis, Seth C. 2014. “Journalism in An Era Of Big Data: Cases, Concepts, and Critiques.” *Digital Journalism*. doi: 10.1080/21670811.2014.976399.

Lewis, Seth C. and Oscar Westlund. 2014. “Big Data and Journalism.” *Digital Journalism*. doi: 10.1080/21670811.2014.976418.

Mair, John, Richard Lance Keeble, Paul Bradshaw, and Teodora Beleaga. 2013. *Data Journalism: Mapping the Future*. Suffolk: Abramis academic publishing.

Mangobug (2012, January 3). Re: What's the Facebook's Graph API call limit? [online forum comment]. Stackoverflow. Accessed December 19, 2014. http://stackoverflow.com/questions/8713241/whats-the-facebooks-graph-api-call-limit.

Netvizz. 2015. “Instructions”. Accessed March 18, 2015. https://wiki.digitalmethods.net/Dmi/ToolNetvizz.

Open Knowledge Foundation. 2012. “Open Data Handbook Documentation. Release 1.0.0.” Accessed December 19, 2014. http://opendatahandbook.org/pdf/OpenDataHandbook.pdf, retrieved 19.12.2014.

Parasie, Sylvain. 2014. “Data-Driven Revelation? Epistemological Tensions in Investigative Journalism in the Age of “Big Data”.” *Digital journalism*. doi:10.1080/21670811.2014.976408.

Semenov, Alexander. 2013. “Principles of Social Media Monitoring and Analysis Software.” PhD diss., University of Jyväskylä. https://jyx.jyu.fi/dspace/bitstream/handle/123456789/41559/978-951-39-5225-9.pdf.

Semenov, Alexander., Jari Veijalainen, and Alexander Boukhanovsky. 2011. “A Generic Architecture for a Social Network Monitoring and Analysis System.” In *The 14th International Conference on Network-Based Information Systems, USA: IEEE Computer Society*, edited by Leonard Barolli, Fatos Xhafa, Makoto Takizawa, 178–185. CA: Los Alamitos. Doi: 10.1109/NBiS.2011.52.

Statistics Finland. 2014. *Use of Information and Communications Technology by Individuals 2014: One half of Finnish residents participate in social network services.* Accessed November 10, 2014. http://www.stat.fi/til/sutivi/2014/sutivi_2014_2014-11-06_tie_001_en.html.

Steensen, Steen and Laura Ahva. 2014. “Theories of Journalism in the Digital Age: An Exploration and Introduction.” *Digital Journalism* 3 (1): 1–18. doi: 10.1080/21670811.2014.927984.

Taloussanomat. December 21, 2013. Fennovoima ja Rosatom sopimukseen ydinvoimalasta [Fennovoima and Rosatom make a deal of nuclear power plant]. Accessed January 1, 2015. http://www.taloussanomat.fi/kotimaa/2013/12/21/fennovoima-ja-rosatom-sopimukseen-ydinvoimalasta/201317748/12.

Ukkonen, Antti. 2013. ”Big Data ja Laskennalliset Menetelmät”. In *Otteita Verkosta. Verkon ja Sosiaalisen Median Tutkimusmenetelmät [Excerpts from the Web: Research Methodology for the Web and Social Media]*, edited by Salla Maaria Laaksonen, Janne Matikainen, and Minttu Tikka, 274–304. Tampere: Vastapaino.

Valio out of Fennovoima’s nuclear plant project [Valio pois Fennovoiman ydinvoimala-hankkeesta]. 2014. Facebook. Accessed December 19, 2014. https://www.facebook.com/pages/Valio-pois-Fennovoiman-ydinvoimala-hankkeesta/175916569145235?ref=ts&fref=ts.

Weber, Matthew S. and Peter Monge. 2011. “The Flow of Digital News in a Network of Sources, Authorities, and Hubs.” *Journal of Communication* 61 (6): 1062–1081. doi:10.1111/j.1460-2466.2011.01596.x.

YLE. March 27, 2014. Fennovoiman ydinvoimalan varma suomalaisosuus putosi alle puoleen [Fennovoima’s secured Finnish finance share dropped under a half]. Accessed January 3, 2015. http://yle.fi/uutiset/fennovoiman_ydinvoimalan_varma_suomalaisosuus_putosi_alle_puoleen/7160093.

Zlatanov, Biser V. and Maya F. Koleva. 2014. “Networks of Collective Power: (Non)Movements and Semantic Networks.” Paper presented at the ECREA 2014 European Communication Conference, Lisbon, November 12-15.

Table 1. Facebook's main concepts and their explanations

Concept	Explanation
API	Defines how a computer system can be accessed by another computer system, what data it can access and on what premises. Facebook has two different APIs, Graph API and FQL (Facebook Query Language), that allow other computer systems to automatically access Facebook. Both APIs allow the access to the same underlying system that runs Facebook (read more from Facebook 2014e; 2014f).
Feed	Central element of Facebook where posts including status updates and links are published. Individual users have their own personal feeds and so do groups, pages, and applications. Another name for a feed is 'stream'. The Graph API uses the term 'feed' whereas FQL uses the term 'stream'.
Post	Individual entry to a feed. It can contain text, image and video content, file, link, users associated with the entry, location and privacy settings. Users who can access the post can share, like and comment the post.
Comment	Entry that can be targeted on most types of content on Facebook. It can contain text, links and photos. Users can like a comment or reply to it.
Like	Action that a user can make to notify the creator of a post or a comment that the user has liked the entry. Everybody who can see the original entry can also see all 'likes'.
Share	Act where a user shares a post generated by someone else on their own feed, or posts it to a feed of a friend, group or page. Information about shares on Facebook is restricted. If a user has permission to read the entry, it is possible to read a list of users who shared the article if their privacy settings allow this.
User	Account of an individual user. Detailed personal information on Facebook is restricted. If the user is not a friend or the user has not chosen to relax their privacy policies, the only information that can be retrieved from the user are first name, last name, gender, age, friend count and subscriber count.

Table 2. Titles and functions of the data warehouse tables

Table name	Table function
f_event	Fact table for storing posts and comments. The table has a parent-child structure, where posts act as parent events for comments.
d_event_type	Dimensional table for describing the event type. The event can be either a post or a comment. In case of posts, also information about the type of post will be saved: app story, event created, link posted, photos posted, post on wall, status update, and video posted.
d_content	Dimensional table for storing content of an event. In case of a post not only text written by the user, but also possible file attachments, web-links and any caption content generated by Facebook are saved. The same is true about comments, although there are no files attached to them.
d_entity	Dimensional table for storing the information about an entity connected to an event. All different Facebook entities such as users, pages, groups and application are handled as entities.
r_like	Relation table between an event and an entity.

Table 3. Key concepts of Facebook and their counterparts on Twitter

Facebook	Twitter	Functionality
Stream	@ or #-tag	Stream of content organized around certain user or specific subject.
Post	Tweet	Content that users send to their social networks.
Comment	Reply	Content that has a link to previous content to which it has follow-up relation.
Share	Retweet	Content that users have forwarded to their networks.
Like	Favorite	Shared sign between users of one user liking a specific content.

Table 4. Excerpt of a data table of weekly posting quantities and posts' related activities

Year	Month	Week	Posts	Posts (source)	Posts (other)	Posters	Shares	Shares (source)	Shares (other)	Likes	Likes (source)	Likes (other)	Comments	Comments (source)	Comments (other)	Comment Likes	Comment Likes (source)	Comment Likes (other)
2013	11	45	13	10	3	3	69	69	0	23	23	0	8	5	3	4	3	1
2013	11	46	18	18	0	0	197	197	0	64	64	0	19	19	0	28	28	0
2013	11	47	5	5	0	0	72	72	0	15	15	0	7	7	0	12	12	0
2013	11	48	5	4	1	1	16	16	0	10	8	2	4	4	0	3	3	0
2013	12	49	11	8	3	2	127	127	0	40	38	2	10	10	0	5	5	0
2013	12	50	11	11	0	0	109	109	0	27	27	0	18	18	0	14	14	0
2013	12	51	11	8	3	1	64	64	0	26	22	4	11	5	6	7	4	3
2013	12	52	5	3	2	2	17	17	0	7	5	2	6	5	1	4	4	0
2013	12	1	12	11	1	1	116	116	0	40	36	4	17	15	2	22	21	1
2014	1	2	4	4	0	0	7	7	0	8	8	0	5	5	0	5	5	0
2014	1	3	8	7	1	1	43	43	0	16	16	0	10	8	2	7	5	2

Table 5. Excerpt of data table with focus on links and their escalated activity

Source	Referred	Referrers	Referrers (user)	Referrers (page)	Referrers (group)	Likes	Shares	Comments	Likes (avg)	Shares (avg)	Comments (avg)
www.facebook.com	222	9	6	3	0	3661	1279	1038	16.49	5.76	4.68
yle.fi	57	8	3	5	0	204	281	49	3.58	8.03	0.86
www.kaleva.fi	40	9	3	6	0	258	270	32	6.45	15	0.8
www.taloussanomat.fi	29	10	2	8	0	86	72	15	2.97	10.29	0.52
www.hs.fi	23	6	0	6	0	84	155	23	3.65	12.92	1
www.adressit.com	21	9	3	6	0	42	40	3	2	2.67	0.14
www.talouselama.fi	20	7	2	5	0	113	63	43	5.65	7	2.15
www.uusisuomi.fi	18	8	2	6	0	59	69	6	3.28	13.8	0.33
globalpost.com	16	1	1	0	0	0	0	0	0	0	0
www.iltasanomat.fi	10	7	1	6	0	9	4	2	0.9	2	0.2
www.tekniikkatalous.fi	9	2	1	2	0	18	13	5	2	1.86	0.56
puhutaanydinvoimasta.fi	6	3	1	3	0	105	125	18	17.5	31.25	3
www.kauppalehti.fi	6	4	1	2	0	6	3	5	1	1	0.83
www.greenpeace.org	5	2	1	1	0	11	1	1	2.2	0.33	0.2
www.mtv.fi	5	1	1	1	0	93	86	8	18.6	21.5	1.6

Figure 1. Event data centric star schema of tables and their connections

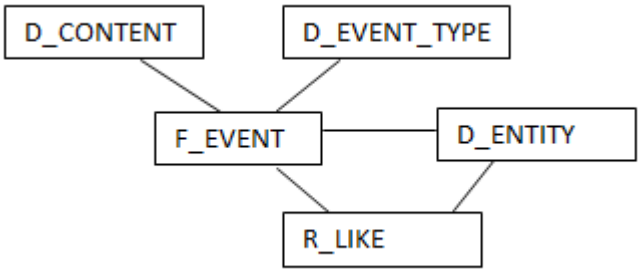

Figure 2. Weekly amounts of posts and tied activities of shares, likes, and comments (2011-2014)

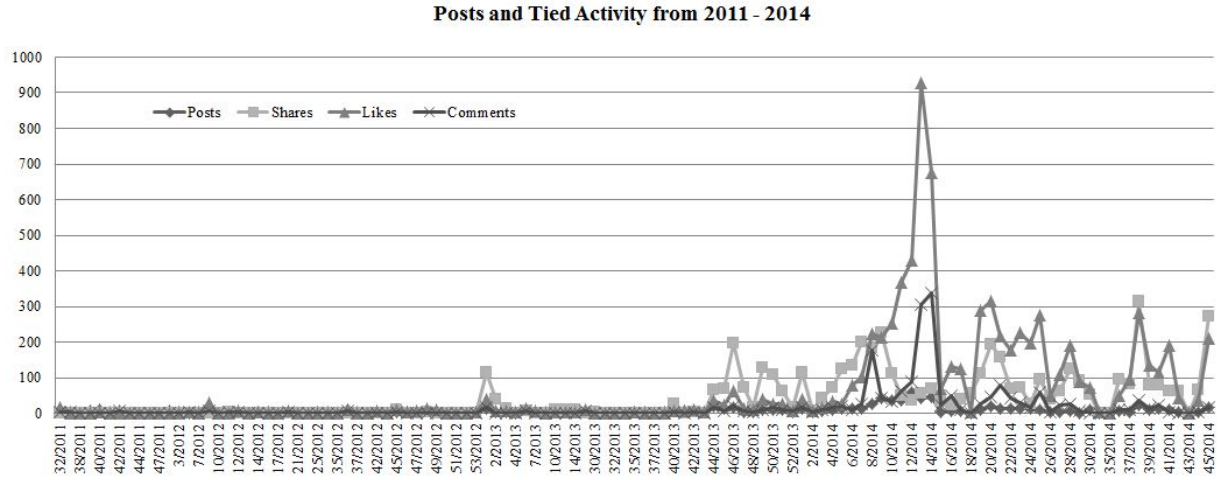

Figure 3. Four most linked news sources and their escalated activity of likes, shares, and comments