Studying Facebook via Data Extraction: The Netvizz Application

Bernhard Rieder
University of Amsterdam
Turfdraagsterpad 9
1012TX Amsterdam
rieder@uva.nl

ABSTRACT
This paper describes Netvizz, a data collection and extraction application that allows researchers to export data in standard file formats from different sections of the Facebook social networking service. Friendship networks, groups, and pages can thus be analyzed quantitatively and qualitatively with regards to demographical, post-demographical, and relational characteristics. The paper provides an overview over analytical directions opened up by the data made available, discusses platform specific aspects of data extraction via the official Application Programming Interface, and briefly engages the difficult ethical considerations attached to this type of research.

Author Keywords
research tool, social networking services, Facebook, data extraction, social network analysis, media studies

ACM Classification Keywords
J.4 Social and Behavioral Sciences

INTRODUCTION

In October 2012, Facebook announced that it had reached the symbolic number of one billion monthly active users. [4] This arguably makes it one of the biggest media organizations in the history of humankind, contested only by Google's collection of services in terms of daily worldwide audience size and engagement. Traditional corporations dwarf these massive Internet companies when it comes to the size of their workforce – Facebook employed a mere 4500 people at the end of 2012 – but the sheer number of "[p]eople [who] use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them" [4] is simply gigantic. It is no wonder, then, that researchers from many areas of the human and social sciences have moved quickly to study the platform: a recent review article [19] identified 412 peer-reviewed research papers that follow empirical approaches, not counting the numerous publications employing conceptual and/or critical approaches. While traditional empirical methods such as interviews, experiments, and observations are widely used, a growing number of studies rely on what the authors call "data crawling", i.e. "gleaning information about users from their profiles without their active participation" [19]. This paper presents a software tool, Netvizz, designed to facilitate this latter approach.

Research methods using software to capture, produce, or repurpose digital data in order to investigate different aspects of the Internet have been used for well over a decade. Datasets can be exploited to analyze complex social and cultural phenomena and *digital methods* [12] have a number of advantages compared to traditional ones: advantages concerning cost, speed, exhaustiveness, detail, and so forth, but also related to the rich contextualization afforded by the close association between data and the properties of the *media* (technologies, platforms, tools, websites, etc.) they are connected with; data crawling necessarily engages these media through the specifics of their technical and functional structure and therefore produces data that can provide detailed views of the systems and the use practices they host. The study of social networking services (SNS) like Facebook, however, introduces a number of challenges and considerations that makes the scholarly investigation of these services, their users, and the various forms of content they hold significantly different from the study of the open Web. This paper discusses some of the possibilities and difficulties with the data crawling approach applied to Facebook and introduces a tool that allows researchers to generate data files in standard formats for different sections of the Facebook social networking service without having to resort to manual collecting or custom programming. I will first introduce some of the approaches to data extraction on SNS, in order to situate the proposed tool. I will then introduce the Netvizz application and provide a number of short examples for the type of analysis it makes possible. Before concluding, I will discuss two further aspects that are particularly relevant to the matter at hand: research via Application Programming Interfaces (API) and the question of privacy and research ethics. While this paper contains technical descriptions, it is written from a media studies perspective and therefore focuses on aspects most relevant to media scholars.

WebSci'13, May 2–4, 2013, Paris, France.

STUDYING FACEBOOK THROUGH DATA EXTRACTION

The study of Internet platforms via data extraction has seen fast growth over the last two decades and the recent excitement around the concept of *big data* seems to have added additional momentum to efforts going into this direction. [9] For researchers from the humanities and social sciences, the possibility to analyze the expressions and behavioral traces from sometimes very large numbers of individuals or groups using these platforms can provide valuable insights into the arrays of meaning and practice that emerge and manifest themselves online. Besides merely shedding light on a "virtual" space, supposedly separate from "real life", the Internet can be considered as "a source of data about society and culture" [12] at large. The promise of producing *observational* data, i.e. data that documents what people do rather than what they say they do, without having to manually protocol behavior, expressions, and interactions is particularly enticing to researchers. SNS in general, and the gigantic Facebook platform in particular, can be likened, on a certain level, to observational devices or even to experimental designs: the "captured" data are closely related to meticulously constructed technical and visual forms – functionalities, interfaces, data structures, and so forth – that function as "grammars of action" [1], enabling and directing activities in distinct ways by providing and circumscribing possibilities for action and expression. Even if the design of this large-scale social experiment is specified neither by nor for social scientists and humanists, the delineated and parametered spaces provided by SNS confer a controlled frame of reference to gathered data. No wonder that Cameron Marlow, one of the research scientists working at Facebook considers the service to be "the world's most powerful instrument for studying human society" [16]. In order to better understand how such data can be gathered, a short overview of existing approaches is indispensable.

Existing Approaches

The already mentioned review paper [19] distinguishes five categories of empirical Facebook research: descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions, and privacy and information disclosure. It is not difficult to see how approaches gathering data from or through the platform can be useful for each of these areas of investigation. The question, then, is what data can actually be accessed and how this is to be done, considering that the particular technique chosen has important repercussions for the scope of what can be realistically acquired.

One can largely distinguish two general orientations when it comes to collecting digital data from SNS through software-based tools: first, researchers can recruit participants, through Facebook itself or from the outside, and gather data by asking them to fill out questionnaires, often via so called Facebook applications[1]. [11] While this method certainly differs from traditional ways of recruiting participants in terms of logistics and sampling procedures, it is not fundamentally different from online surveying in general.[2] Second, data can be retrieved in various ways from the pools of information that the Facebook platform *already* collects as part of its general operation. This latter approach, which is the focus of this paper, is fueled by data derived from both sides of the distinction Schäfer makes between "implicit and explicit participation" [14], referring to the difference between information and content deliberately provided by users, e.g. by filling out their profiles, and the data collected and produced by logging users' actions in sometimes minute detail. While Facebook members share content, write messages, and curate their profiles, they also click, watch, read, navigate, and so forth, thereby providing additional data points that are stored and analyzed. Because these activities revolve around elements that have cultural significance – liking a page of a political party is more than "clicking" – these data are not simply behavioral, but allow for deeper probing into *culture*. For research scholars, there are three ways by which to gain access to these data, with significant differences between approaches in terms of technical requirements and institutional positioning:

Direct database access to the company's servers is reserved to in-house researchers or cooperation between a SNS and a research institution. [17] Certain companies also make data "donations", for example Twitter deciding to transfer its complete archive to the Library of Congress, albeit with a significant delay. The data made accessible in these ways are generally very large and well structured, but often anonymized or aggregated. Partnering with a platform owner is certainly the only (legal) way to gain access to *all* collected data, at least in theory.

Access through sanctioned APIs makes use of the machine interfaces provided by many Web 2.0 services to third-party developers with the objective of stimulating application development and integration with other services in order to provide additional functionality and utility to users. These interfaces also provide well-structured data, but are generally limited in terms of which data, how much data, and how often data can be retrieved. Conditions can vary significantly between services: in contrast to Twitter, for example, Facebook is quite restrictive in terms of what data can be accessed, but imposes few limits on request frequency. Companies also retain the right to modify or close their data interfaces, which can lead to substantial problems for researchers.

---

[1] A Facebook application is a program that is provided by a third-party but integrates directly into the platform.

[2] One should note that studies using questionnaires on Facebook often access profile data as well.

User interface crawling can be done manually, but usually employs so-called *bots* or *spiders* that read the HTML documents used to provide graphical interfaces to users, either directly at the HTTP protocol level or via browser automation from the rendered DOM.[3] [8] These techniques can circumvent the limitations of APIs, but often at the price of technical and legal uncertainties if a platform provider's permission is not explicitly granted. In the case of Facebook, bot detection mechanisms are in place and suspicious activity can quickly lead to account suspension.

If performed on a large scale, all of these approaches require either custom programming or considerable amounts of manual work. The focus points and requirements for research and teaching do, however, bear marks of resemblance and Facebook itself is designed around a limited number of functionalities or "spaces". One can therefore argue that general-purpose tools may be envisioned that provide utility to a variety of research projects and interests. Several such *data extractors* targeting Facebook have been developed over the last years, invariably using sanctioned APIs for data gathering. These tools generally export data in common formats and they focus on specific sections of the platform – partly by choice, partly due to limitations imposed by the platform itself. Their goals are also similar: to lower the technical and logistical requirements for empirical research via data analysis in order to further the ability of researchers to study a medium that unites over a billion users in a system that is essentially conceived as a walled garden. In what follows, I describe the Netvizz application[4], a tool designed to help research scholars in extracting data from Facebook.

Similar Work

The enormous success of Facebook has prompted the emergence of a large number of analytics tools for marketing purposes, which often focus on *pages*, the section of Facebook that brand communication and consumer relations rely on, due to their public showcase character. Because these tools are generally built for monitoring marketing campaigns, they target page *owners* rather than researchers interested in studying a page. For this reasons – and the sheer number of tools available – I will leave these applications to the side.

There are, however, two tools that function as general-purpose data extractors for researchers studying Facebook. *NameGenWeb*[5] originated at the Oxford Internet Institute and provides the possibility of exporting a user's friendship network, i.e. all of the user's friends, the friendship connections between them, and a wide array of variables for each user account extracted. Another application, the *Social Network Importer*[6], a plug-in for the *NodeXL* network analysis and visualization toolkit developed by an international group of scholars, provides similar functionality for downloading personal networks, but also a means to extract extensive data from Facebook pages, including monopartite[7] networks for users and posts, based on co-like or co-comment activities, and bipartite networks combining the two in a single graph. One should also mention Wolfram Alpha's "Facebook report"[8] in this context: while it does not make raw data available, and therefore limits in-depth analytics using statistical or graph theoretical approaches, the tool provides a large number of analytical views on personal networks.

The Netvizz application provides "raw" data for both personal networks and pages, but provides data *perspectives* not available in other tools, e.g. comment text extraction; it also provides data for groups, a third functional space on Facebook. Running as a Web application, Netvizz does not require the use of Microsoft Excel on Windows like *NodeXL* and thereby further lowers the threshold to engagement with Facebook's rich data pools. The next section will introduce the application and its different data outputs in more detail.

THE NETVIZZ APPLICATION

The Netvizz application was initially developed by the author in 2009 as a practical attempt to study Facebook's API as a new media object in its own right[9] and to gauge the potential of using natively digital methods [12] to study SNS. Because of the positive reactions and high uptake, the application was developed into a veritable data extractor that provides outputs for different sections of Facebook in standard formats.[10] Before introducing the different

[3] The latter approach has become more common due to the fact that sites are increasingly using programming languages (mostly JavaScript) to assemble pages client-side rather than sending finished documents described in a markup language (mostly HTML).

[4] https://apps.facebook.com/netvizz/

[5] https://apps.facebook.com/namegenweb/

[6] http://socialnetimporter.codeplex.com/

[7] Monopartite graphs contain nodes that are all of the same kind (e.g. users). Bipartite graphs include two types of nodes (e.g. users and posts), and so forth.

[8] http://www.wolframalpha.com/facebook/

[9] APIs as *objects* of research for new media scholars are only slowly coming into view, despite their importance for the Web as data ecosystem. A separate publication will detail empirical approaches to studying APIs from a critical media studies perspective.

[10] Data formats were chosen for their generality and simplicity. Network outputs use the GDF format introduced with the GUESS graph analysis toolkit. Tabular outputs use a simple tab separated format that can be opened in virtually all spreadsheet applications and statistical packages.

features, it is necessary to briefly discuss the Facebook API and those characteristics that are relevant to research procedures and data quality.

Data Access via the Facebook API

As indicated above, Netvizz is a simple Facebook application written in PHP that runs on a server provided by the Digital Methods Initiative[11]. It is part of Facebook's app directory and can be found by typing the name into the platform's main search box. Like any other Facebook application, it requires users to log in with an existing Facebook account to be able to access any data at all.

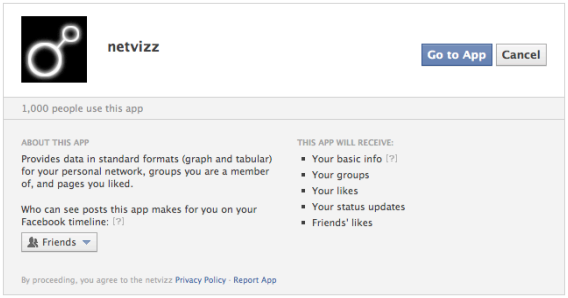

Figure 1. The Netvizz app permission request page.

A vast SNS that deals with intimate and potentially sensitive matters is likely to implement rather strict privacy policies and this is – to a certain extent – also the case with Facebook. The construction of the Facebook API reflects these concerns in at last four ways that are significant here:

First, every probe into the data pool is "signed" with the credentials of a Facebook user whose actual status on the platform defines the scope of which data can be accessed. For example, detailed user data can generally only be extracted from accounts a user is friends with and one has to be a member of a group to extract any data from it.

Second, users' privacy settings play a role in what data can be exported. If one user excludes another from seeing certain elements on his or her profile, an application operating with the latter's credentials will also be blocked from accessing those elements.

Third, every application is required to explicitly ask for permission to access different data elements.[12] These requests are displayed to the user when she first uses the application. Figure 1 shows the permission dialogue for the Netvizz application. While these permissions have to be given for the application to work, users can limit the data made available to applications used by their friends in their preferences.

Fourth, certain elements that are visible on the level of the user interface are not available through the API. The user view count displayed on each post in a group, for example, is (currently) not retrievable and certain data elements, such as friends' email addresses, are equally off limits by design.

While we can expect scholars using the Netvizz application to grant all the permissions[13] it asks for – it will simply not work otherwise – users' privacy settings are indeed relevant when it comes to interpreting the retrieved data: from a technical perspective, it is not possible to know whether an empty field is empty because the user has not filled in the specific data or because the privacy settings prohibit access. This must be taken into account when making assumptions on the basis of missing data. User profile data, in particular, should be handled with prudence. Other data, such as page engagement and friendship connections in personal networks and groups, can be considered robust, however.

Overview

The Netvizz application currently extracts data from three different sections of the Facebook platform:

Personal networks are considered in two different ways. First, the friendship network feature provides a simple undirected graph file where the friends of the logged user are nodes and friendship connections edges. Sex, interface language, and a ranking based on the account creation date[14] are provided for each user and counts for posts and likes can be requested as an option. Friendship networks often cluster around significant places in a user's life, e.g. geographies or institutions such as high school, university, workplaces, clubs, and so forth. Second, a bipartite "like network" can be generated that formalizes both users and liked entities (all elements already represented in Facebook's Open Graph[15] are extracted) as nodes, a user liking a page generating an edge. This network, examined via a graph analysis toolkit, will arrange both users and liked objects around cultural affinity patterns, foregrounding *post-demographic* [13] variables.

Groups can be explored in a similar fashion as friendship networks, although the API currently limits the number of users one can retrieve from a group to 5000. For larger groups, a random subset of users is provided. A second

[11] https://www.digitalmethods.net

[12] For details concerning the permission structure refer to: http://developers.facebook.com/docs/reference/login/

[13] The Netvizz application does not store or aggregate any of the extracted data in a database and the generated files are deleted in regular intervals.

[14] The unique identifiers for accounts on Facebook are numbered consecutively, which means that the lower the number, the older the account. Netvizz simply adds a ranking to the output that orders accounts by their age.

[15] For more information on how Facebook represents entities in the *Open Graph* concept modeling system, see: https://developers.facebook.com/docs/concepts/opengraph/.

feature also provides a *social* graph, but one that is based on interactions between group members through the posts sent to a group. If one user likes or comments on another user's post, a directed edge between the two users is created, each interaction adding weight to the edge.

Pages are represented as a bipartite network, with both posts (up to 999 latest posts) and users as nodes. If a user comments on or likes a post, a directed edge between user and post is created. This way, one can not only detect the most active users, but also identify the posts that produced the highest amount of engagement. The latter data are also provided in a tabular data file, ready for statistical analysis. To make content analysis easier, a third file containing user comments, grouped per post, is generated. The application allows selecting whether posts made by users should be included, in addition to posts made by the page owner.

ANALYTICAL DIRECTIONS

The two types of data files provided by Netvizz – network files and tabular files – already indicate basic directions for analytical approaches, the former allowing for the application of concepts and methods from Social Network Analysis [15] and Network Science [18], while the latter points towards more traditional statistical techniques. Before describing analytical approaches in more detail, a short comment on modes of analysis – and in particular visualization – is in order.

Analysis and Visualization

One of the reasons for choosing simple and common file formats for outputs in Netvizz was the need to compensate for the lack of an actual visual and analytical interface in the application itself. There are, indeed, a number of Facebook applications available that produce direct visual representations, generally of personal networks, which greatly facilitates the initial encounter with the data in question for researchers with little or no training in quantitative research. Because these tools are mostly visualization widgets that do not target researchers and offer little to no analytical methodology beyond the visual display itself, one of the initial intentions was to design Netvizz as a bridge between Facebook data and the various network analysis toolkits available today, such as GUESS[16], Pajek[17] or the very easy to use Gephi[18]. The last program, in particular, must be credited with significant lowering the threshold to working with network analysis and visualization. Netvizz voluntarily inscribes itself in a movement, epitomized by tools such as gephi and the work of the Amsterdam-based Digital Methods Initiative[19] and other groups, that aims at bringing data-driven analysis to a wider audience and, specifically, to an audience that includes those regions of the social sciences and humanities that have been shunning quantitative and computational methods because of the epistemological and methodological commitments often associated with quantification and formalization. Lowering the threshold to using computer-based analytical methods is therefore not simply a service to long-time practitioners, but an attempt to see in what way and how far these methods can be useful in contexts where the dominant "styles of reasoning" [7] are based on interpretation, argumentation, and speculation, and build on conceptualizations of human beings and their practices that simply cannot be formalized as easily as theoretical frameworks like behaviorism or social exchange theory.

In this context, visualization has been presented as a means to profit from the analytical capacities afforded by software without having to invest years into the acquisition of skills in statistics or graph theory. While the data provided by Netvizz can certainly be used to calculate correlation coefficients as well as network metrics, focus was put on facilitating analysis through visualization. There is, however, no need to juxtaposition mathematical and visual forms of analysis; as Figure 2 demonstrates, the latter can not only help in communicating the results provided by the former, but adds a way of relating to the data that can provide a significant epistemic surplus.

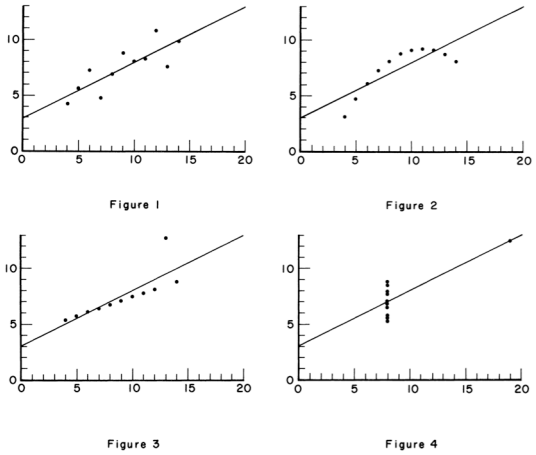

Figure 2. Four scatter-plots from [2]. They have identical values for number of observations, mean of the x's, mean of the y's, regression coefficient of y on x, equation of regression line, sum of squares of x, regression sum of squares, residual sum of squares of y, estimated standard error of bi, and multiple r2. Yet, the differences between the dataset are strikingly obvious to our eyes. Anscombe uses this example to make an argument for the usefulness of visualization in statistics beyond the communication to a larger audience.

Independently of its application to actual empirical analysis of Facebook data, Netvizz should thus be considered a pedagogical tool that can help in getting started with quantitative methodology, network analysis, and the

[16] http://graphexploration.cond.org

[17] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[18] https://gephi.org

[19] https://digitalmethods.net

required software. While one could argue that network visualizations are images and therefore intuitively accessible and "readable", there are also arguments that point into the opposite direction. It is easy to show how different graph layout algorithms highlight particular properties of a network and familiarity with a dataset can go far in helping novice users understand what is actually happening when they use software to work with graph data. Because many people are intimately familiar with their Facebook networks, they can more easily see what the software does, and what kind of epistemic surplus one can potentially derive from network analysis.

Analytical Perspectives

In actual research settings, Netvizz can provide data relevant to many different approaches and research questions. One can also consider different embeddings in the logistics of research projects: it is imaginable that a study recruits users to investigate patterns in social relations, but instead of asking them for access to their accounts, they encourage them to run the Netvizz application from their profile and share the data with the researchers. Descriptive approaches to user profiling could thus complement traditional socio-economic descriptors with *post-demographic* properties [13] in the form of like data and the *relational* data represented by friendship networks. It is worth mentioning that Netvizz uses the unique Facebook account identifiers as "keys" for nodes in the GDF format; this means that all network files can be combined to form larger networks because the same user appearing in two different files will be a single node if the networks are combined, e.g. in gephi.

The group and page features also enable or facilitate data-driven approaches to studying Facebook users and uses without requiring access to individual accounts. In the case of groups, one needs to be a member to access its data; in the case of pages, liking it is enough to make it show up in the Netvizz interface. The analytical possibilities afforded by the second perspective are explored in more detail via two short case studies in the following section, but one could classify analytical dimensions along a series of very basic questions:

Who? This concerns studies of users (profile data), their relations (friendship patterns and interactions), and the larger social spaces emerging through groups and pages.

What? For personal networks, this relates mainly to *likes*, while pages allow for an investigation into *posts*, in particular concerning media types and audience engagement.

Where? For all outputs containing information about users, interface language is provided in a comprehensive way, because users do not have the possibility to prevent applications from receiving this information. While interface language is certainly not a perfect stand-in for locality, it allows engaging the question of geography in interesting ways.

When? Temporal data is limited to pages, but here, a timestamp for each post and comment is provided, allowing for investigating page and user activity over time.

EXAMPLES

To make the provided directions for analysis more tangible, this section briefly outlines two case studies investigating the use of Facebook in political activism online, more precisely its use by the anti-Islam movements that have grown at a rapid pace, in particular since the 9/11 attacks. The first example focuses on a group and the second on a page. Both examples mobilize concepts and techniques from Social Network Analysis (SNA), which developed out of the work of social psychologists Jacob Moreno and Kurt Lewin in the 1930s and 1940s. Although its tight relationship with social exchange theory [3] has granted a certain amount of visibility to SNA, it is only the wide availability of relational data and the software tools to analyze these data that the approach has gained the popularity it enjoys today. The main tenant of SNA is to envision groups and other social units as *networks*, that is, as connected ensembles that emerge from tangible and direct connections (friendships, work relationships, joint leisure, direct interactions, etc.) rather than as social *categories* that are constructed on the bases of shared (socio-economic) properties instead of actual interactions. This approach is particularly promising when applied to Facebook groups.

The "Islam is Dangerous" Group

The "Islam is Dangerous" group is an "open" group on Facebook, which means that its shared posts and members are visible to every other Facebook user. At the time of writing, the group had 2339 members and was mainly dedicated to sharing information about atrocities, crimes, infractions or simply deviations from cultural standards by Muslims.

A first approach used Netvizz for extracting all friendship connections between all the members of the group. While it is difficult to imagine an "average" Facebook group, a first finding is constituted by what seems to be a relatively high network density of 0.019. An average degree of 39.7 is a second indicator that this is group hosts a tightly knit collective rather than a loosely associated group merely sharing information on a subject. Friendship patterns are, however, not evenly distributed. While 18.3% of the group members have no friendship connection with other members – a population attracted by the subject matter rather than through social contacts? – 37.2% have at least 20 connections and 14.8% 100 or more.

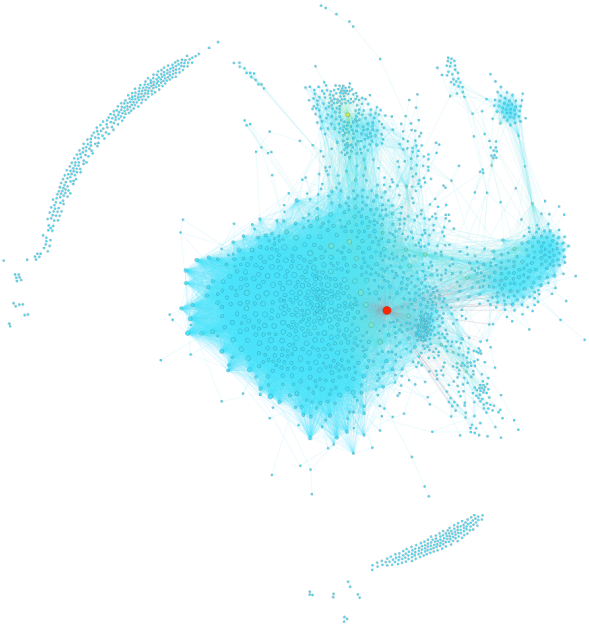

Figure 3. Friendship graph for the "Islam is Dangerous" group, colors represent betweenness centrality via a heat scale (blue => yellow => red).

While counting connections may be one way to identify leaders in a group, network analysis provides an extensive arsenal of techniques to analyze graphs in more specific ways. Figure 3 shows a spatialized visualization of the group (using gephi) and points to our ability to use advanced graph metrics to further analyzed the dataset by coloring nodes with a metric called *betweenness centrality*. This measure expresses a node's positioning in the larger topology of a graph and it can be very useful for detecting strategic positioning rather than popularity or social status. A person having high betweenness centrality is considered to be able to "influence the group by withholding or distorting information in transmission" [5] because he or she is located as a passage point between different sections of a network. While there are caveats to consider, betweenness centrality can be likened to Robert Putnam's concept of "bridging" social capital [10], which denotes the capacity to connect separate groups. In our case, this metric identifies the group administrator as the central *bridger*, which points to a group structure that, despite its high connectivity, is held together by a central figure.

The application of betweenness centrality can be seen as an example – a large number of techniques are now available to investigate structure, demarcate subgroups or qualify users in terms of their position in the network. Graph analysis software generally provides implementations of these metrics to researchers.

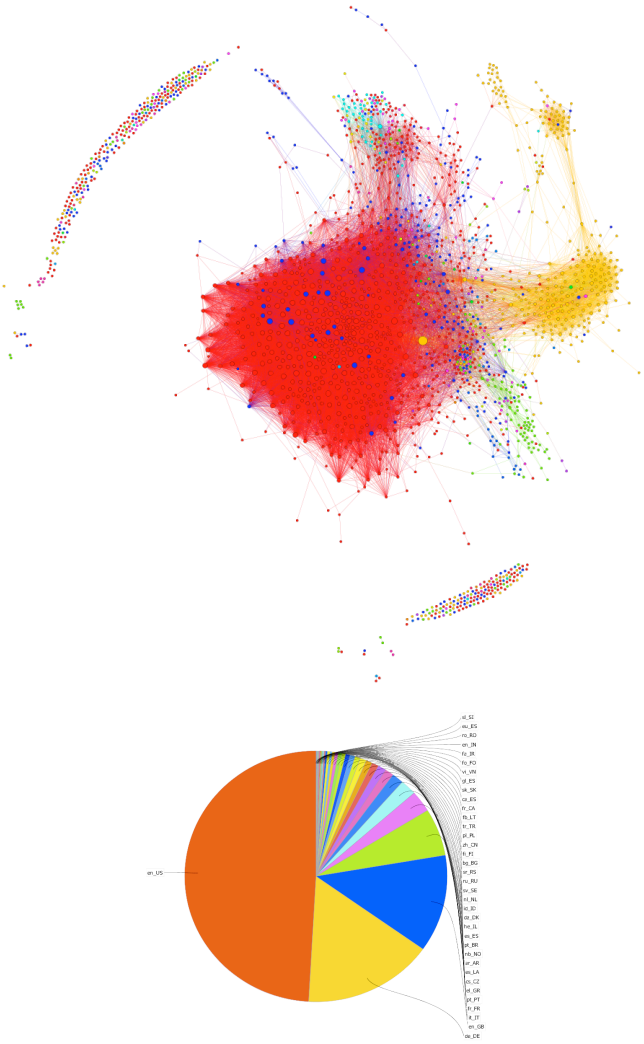

Figure 4. Friendship graph for the "Islam is Dangerous" group, colors represent "locale", i.e. the language of the Facebook interface for a given user.

Another example for types of analysis makes use of the users' interface language ("locale"), one of the few data points available for every Facebook member. Figure 4 shows the same network diagram as above, but uses locale to color nodes. We can see that there is a densely connected cluster of English speakers (both US and UK) that dominates the group, but smaller subcommunities, in particular a German one in yellow, can be identified as well. We can make the argument that this group, despite its high level of connectivity retains a degree of national coherence.

The "Educate children about the evils of islam" page

The second example quickly analyzes the Facebook page entitled "Educate children about the evils of Islam", which had been liked by 1586 users at the time of writing.

When extracting data from pages, Netvizz essentially operates by iterating over the last n (< 999) posts, collecting

the posts themselves, as well as all of the users that like and comment on them. These data can be analyzed in various ways, either as bipartite network (Figure 5) or in more traditional form trough statistical analysis (Figures 6 and 7).

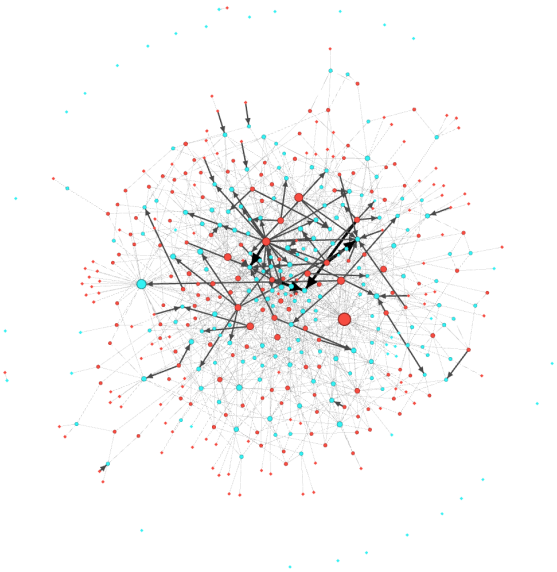

Figure 5. A network diagram showing the last 200 posts (turquoise), as well as the 253 users (red) liking and commenting them.

Network analysis maps interactions on a structural level and allows for the quick identification of particularly successful posts (in terms of engagement) and particularly active users. In this case, what emerges is a picture of a rather lively and intense conversational setting, with a core of loyal visitors that comment and react regularly.

Analyzing the posts over time (Figure 6), we can see that the 200 posts cover a period of less than four weeks, which indicates a high level of investment by the page owner, the only person allowed to post on the page.

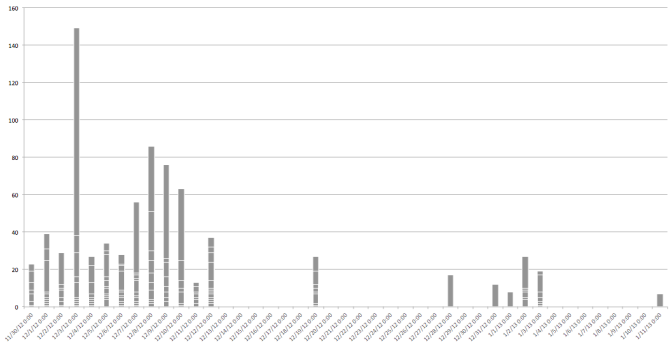

Figure 6. A stacked barchart showing the last 200 posts according to the days they were posted on; values indicate user engagement.

Because Facebook segments posts in content categories, we can also analyze content types, e.g. in relation to how particular types succeed in engaging users.

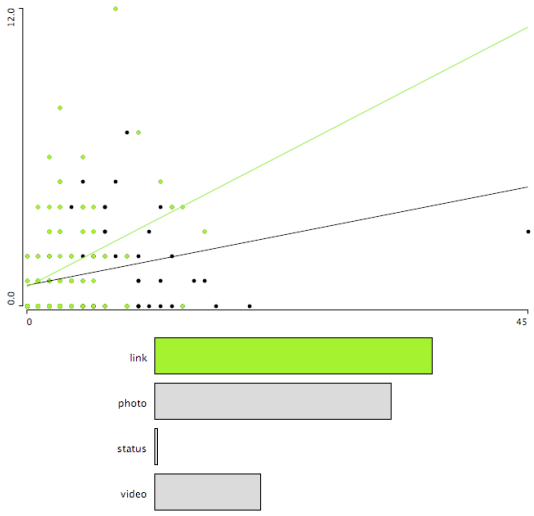

Figure 7. Visualization (using Mondrian) of the content types of the last 200 posts and how often they were liked (x-axis) and commented on (y-axis). Links are highlighted.

Figure 7 shows not only the distribution of content types over the last 200 posts (barchart), but also allows us to correlate these types to user activities. We can learn that links have a higher probability to receive comments, while photos are particularly likely to be liked.

These examples are mere illustrations of the analytical potential the in-depth data Facebook collects and Netvizz extracts. Many other types of analysis – from statistics to content analysis – are possible.

PRIVACY AND RESEARCH ETHICS CONSIDERATIONS

This final sections briefly sketches two aspects related to questions of privacy and research ethics, which would, however, merit a much more in-depth discussion that the space constraints allow.

The Facebook API as privacy challenge

Before discussing ethical considerations of data extraction on Facebook, it is useful to point out that part of the motivation for developing the Netvizz application was an exploration of the Facebook API itself, including the question how it governs access to data and what this means for users' capacity to limit or curate the way their data is accessible to others. This question is important because machine access needs to be treated differently than user interface access to data. While the latter is generally put to the front, the former allows for much more systematic forms of high speed and high volume data gleaning. Manual surveillance of activity is certainly possible, but I would argue that the largest part of user data collection by third parties on Facebook is performed via software that uses similar technological strategies as the Netvizz application. The application – and the knowledge gained by developing it – should therefore also be considered as an indicator of the types of information that other Facebook applications

can get access to and certainly make extensive use of. While the fine-grained permission model holds the promise to limit third party access by asking users explicitly for permission, there is often no possibility for users to actually modulate which rights are granted: the application has to ask for detailed permissions for individual elements, but we can only acquiesce to all request or not use the platform. Access can be revoked *after* installation, but this means that applications can read that data at least once.

As Netvizz shows, a user granting rights to an application generally means that considerable access is given not only to her data, but also to *other* users' data. Application programming for research proposes is useful because of the analytical outcomes it produces or helps to produce, but it should also be considered as an investigation into the technological structures of platforms, which are as relevant to matters of privacy and beyond as they are understudied.

Research ethics

Social scientists have been confronted with the ethical dimension of empirical research well before the advent of the Internet. At no point have answers been easy or clear-cut. Recent debates amongst Internet researchers [20] have tended to put emphasis on the question of individual privacy. We should, however, note that there are significant cultural and political variations when it comes to arguing research ethics. Following Fuchs' critique [6] of the one-sided emphasis on a narrow definition of privacy, I would like to argue that research ethics navigate in a field defined by a number of tensions and competition between different ideals. Putting individuals' privacy on the top of the pyramid is a choice that can be traced to liberal sources of normative reasoning in particular, but we should not forget that these value sources are contingent and culturally colored. Competing ideals, such as the independence of research, larger social utility or the struggle against the encroaching of the private domain on publicness can equally be connected to established traditions in ethical reasoning.

It is clear that national traditions respond to these matters in different ways. While research ethics boards have become the norm in English-speaking countries, such an institutional governance of ethical decisions is hard to imagine in continental European countries such as France, where normative reasoning is concentrated both on the levels of the state and the individual, but only to a lesser degree on the layers in between. Similarly, the study of political extremism, and of the groups and individuals active in such movements, will not be framed in the same way in Germany and the United States, for obvious historical reasons.

What does that mean for Netvizz? Two decisions have been made: first, to anonymize all users for both groups and pages, simply because the number of accounts that can be collected this way is very large. For bigger pages, it is easy to quickly collect data for tens or even hundreds of thousands of user accounts. Second, Netvizz provides an option to anonymize accounts for personal networks. In this case, the complicated weighing of values and research ethics stays in the realm of the user/researcher and are only partially delegated to the programmer.

CONCLUSIONS

This paper has described the Netvizz application, a general-purpose data-extractor for different subsections of the Facebook platform. With a focus on questions relevant to media scholars, in particular, I have contextualized the application in a wider set of research concerns. With Facebook now counting over one billion active users, it is becoming urgent to develop and solidify research approaches to a service, largely constructed as a *walled garden*, that is part of an ongoing privatization of communication, both in terms of economics and accessibility. While there are important limits to what can be done without having to enter into a partnership with the company, the Netvizz application shows that certain parts of Facebook *are* amendable to empirical analysis, after all.

As Netvizz is continuously developed further, additional features will be added in the future. Providing more in-depth data on temporal aspects of user engagement with contents will certainly be one of the next steps.

ACKNOWLEDGMENTS

I would like to thank the attendants of the *Digital Methods Winter School 2013* for their useful comments, in particular Erik Borra and Jean-Christophe Plantin, as well as four anonymous reviewers.

REFERENCES

1. Agre, P.E. Surveillance and Capture: Two Models of Privacy. *The Information Society 10*, 2 (1994), 101-127.
2. Anscombe, F.J. Graphs in Statistical Analysis. *The American Statistician 27*, 1 (1973), 17-21.
3. Emerson, R.M. Social Exchange Theory. *Annual Review of Sociology 2*, (1976), 335-362.
4. Facebook Key Facts. http://newsroom.fb.com/Key-Facts.
5. Freeman, L.C. Centrality in Social Networks. Conceptual Clarification. *Social Networks 1*, 3 (1979), 215-239.
6. Fuchs, C. An Alternative View of Privacy on Facebook. *Information 2*, 4 (2011), 140-165.
7. Hacking, I. *Historical Ontology*. Harvard University Press, Cambridge, MA, USA, 2004.
8. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. and Christakis, N. Tastes, Ties, and Time: a New Social Network Dataset Using

Facebook.com. *Social Networks 30*, 4 (2008), 330-342.

9. Manovich, L. Trending: the Promises and the Challenges of Big Social Data. In Gold, M. *Debates in the Digital Humanities*. The University of Minnesota Press, Minneapolis, MN, USA, 2012, 460-475.
10. Putnam, R.D. *Bowling Alone: the Collapse and Revival of American Community*. Simon and Shuster, New York City, USA, 2000.
11. Quercia, D., Lambiotte, R., Kosinski, M., Stillwell, D., and Crowcroft, J. The Personality of Popular Facebook Users. In *Proc. CSCW 2012*, ACM Press (2012), 955-964.
12. Rogers, R. *The End of the Virtual*. Amsterdam University Press, Amsterdam, The Netherlands, 2009.
13. Rogers, R. Post-Democraphic Machines. In Dekker, A., Wolfsberger, A. *Walled Garden*. Virtual Platform, Amsterdam, The Netherlands, 2009, 29-39.
14. Schäfer, M.T. *Bastard Culture! How User Participation Transforms Cultural Production*. Amsterdam University Press, Amsterdam, The Netherlands, 2011.
15. Scott, J. Social Network Analysis. *Sociology 22*, 1 (1988), 109-127.
16. Simonite, T. What Facebook Knows. *MIT Technology Review*, June 13, 2012.
17. Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. The Anatomy of the Facebook Social Graph. *eprint arXiv:1111.4503*, 2011.
18. Watts, D.J. The 'New' Science of Networks. *Annual Review of Sociology 30*, 1 (2004), 243-270.
19. Wilson, R.E., Gosling, S.D. and Graham, L.T. A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science 7*, 3 (2012), 203-220.
20. Zimmer, M. 'But the Data Is Already Public': on the Ethics of Research in Facebook. *Ethics and Information Technology 12*, 4 (2010), 313-325.