Facebook and its Disappearing Posts: Data Collection Approaches on Fan-Pages for Social Scientists

Erick Behar Villegas

Abstract

Facebook fan-pages are channels of institutional self-representation that allow organizations to post content to virtual audiences. Occasionally, posts seem to disappear from fan-pages, puzzling page administrators and posing reliability risks for social scientists who collect fan-page data. This paper compares three approaches to data collection (manual real-time, manual retrospective, and automatic via NVIVO 10®) in order to explore the different frequencies of posts collected from six institutional fan-pages. While manual real-time collection shows the highest frequency of posts, it is time consuming and subject to man-

Erick Behar Villegas is a Ph.D. candidate at the Institut für Interkulturelle Kommunikation at Ludwig Maximilians Universität München. Correspondence can be directed to erick.behar@hotmail.com.

ual mistakes. Manual retrospective collection is only effective when filters are activated and pages do not show high posting frequency. Automatic collection seems to be the most efficient path, provided the software be run frequently. Results also indicate that the higher the posting frequency is, the less reliable retrospective data collection becomes. The study concludes by recommending social scientists to user either real-time manual collection, or to run a software as frequently as possible in order to avoid biased results by 'missing' posts.

Facebook permeates the lives of millions of people on a daily basis. Some years ago, studies about Facebook mentioned a (then) impressive mark of more than 250 million users (Gjoka, Kurant, Butts, & Markopoulou, 2010). Today, the number of active users easily surpasses 890 million (Facebook, 2015). The complexity that underlies this vast world of virtual interactions may yield valuable data for researchers interested in personal profiles, fan-pages, events, apps, etc. Relevant for this study are fan pages, which depict how organizations and other agents wish to portray themselves to users. These pages work as identity mechanisms (Milolidakis, Akoumianakis, Kimble, & Karadimitriou, 2014, p. 902) that symbolize the marketed idealization of brands, agencies, famous people, and institutions.

The data posted on the timeline of these pages is valuable for researchers using content analysis to extract information. For example, take the evolution of a certain brand and its user feedback as expressed by likes and comments. But what happens if this fan-page data does not

always portray the same content? What happens if posts begin disappearing or if they are filtered through unknown algorithms beyond the page administrator's reach? The latter does not refer to geographically limited posts that appear to users from certain places, but to posts that are not limited by the administrator's own criteria.

Scrolling down the News Feed, a Facebook feature, is one of the most important user activities, yet fan-pages also show high levels of traffic (Facebook, 2014). The older the content is, the less visible it will tend to be for users. In the former Facebook algorithm EdgeRank, time decay was an important factor explaining the loss of a post's relevance (Crossfield, 2013). As new posts flow into fan-page timelines, users must actively access the page and scroll down to see its posts. Users may not even realize or mind that posts are not there anymore. Yet if a social scientist is collecting posts directly from a fan-page, the eventual loss of information may bias his or her study considerably. This paper intends to shed light on eventual data loss by comparing three collection approaches. The final objective is to raise awareness and concretely document the unstable nature of Facebook content. The study also attempts to shed light on the question whether it is better to collect data manually or automatically via software.

Related Literature and Study Insights

Facebook's complexity is reflected by the variety of publications it has inspired. Wilson, Gosling, and Graham (2012) screened several scholarly databases and classified more than 400 Facebook-related articles into five groups: descriptive analysis of users, motivation for using Facebook, identity presentation, the role of Facebook in social

interactions, and information disclosure.

Categories beyond the social sciences can be also traced to computer-science contributions. Studies dealing with data-extraction (aka collection, retrieval), social network structures, graph analysis, *inter alia*, are also available (Bechman & Vahlstrup, 2013; Gjoka et al., 2010; Gjoka, Sirivianos, Markopoulou, & Yang, 2008; Nasution & Noah, 2012; Rieder, 2013; Traud, Kelsic, & Mucha, 2009; Viswanath, Mislove, Cha, & Gummadi 2009; Wilson et al., 2012, pp. 214, 215).

Three examples of this trend are a Google-hosted project named *data-extraction-facebook* (Data Extraction Facebook [DEF], 2009), which analyzes patterns within group memberships, the Center for Ultra-scale Computing and Information Engineering [CUCIS] (2011) project on Social Media, and the Digital Methods Initiative that offers several Facebook-related tools (DMI, 2015). Catanese, De Meo, Ferrara, Fiumara, & Provetti (2012) provide a generous literature review of technical contributions related to Facebook data retrieval while Manning, Raghavan, & Schütze (2009) offer a practical introduction on information retrieval in general.

In their study on Online Social Network modeling, Cormode, Krishnamurthy, and Willinger (2010) integrate the temporal dimension as a complementary element of the usual models that use nodes and edges to model social networks. In their Entity Interaction Networks model, they claim that "it is vital to include temporal information about activities" (p.1). When referring to Facebook, they add that "the type and frequency of […] interactions are much more nuanced than simply recording that two individuals once indicated mutual friendship" (p. 2), pointing

to the necessity of studies that go beyond ties between users. Finally, the authors also mention the importance of considering the lifetime of Facebook lifetime in social media research. These points are key when considering that the present study depends on the variable of time. The findings of this study confirm that time conditions the frequency of fan-page posts, and that their lifetime as 'visible' posts is limited.

It is important to clarify that the issue in question is not the disappearance of pages, user profiles or user posts, but that of single fan-page generated posts. While the subject is relevant for researchers who may lose data due to post disappearance, it seems to be unimportant or irrelevant in studies based on data crawling. For example, Viswanath et al. (2009) addressed user interaction on Facebook referring to the New Orleans Network and crawled user's wall posts. They carried out their second crawl in a period of 3 days and then downloaded the users' wall history (Section 2.1, para. 3). The question whether posts may disappear is not considered.

In another study, Gjoka et al. (2008) refer to this problem, yet indirectly. By screening Facebook analytics, they compared the daily statistics provided by *Adonomics* with the ones reported by Facebook on its application directory in order to test data reliability (p. 32). Although their study does not concern Facebook fan-page posts but user profiles and networks, the importance of "daily statistics" is tacitly linked to data reliability.

This issue of disappearing posts seems to have been addressed explicitly only in Facebook forums and in few press articles. The first source reveals that disappearing posts may be linked to a 'posting mistake'. The user posts

not as the page administrator but as the single user, resulting in missing posts (Facebook Help Forum [FHF], 2015). The second source, press articles, digs deeper into the issue. As reported in *The Mercury* (Otto, 2013), its staff was puzzled by disappearing Facebook posts. The staff commented:

> It seems to be kind of random (…) Some of the links we'll post will stay (…) It was regardless of whether people commented or not, perhaps we don't talk about this, because no one looks for the problem (…) We hadn't really experienced that before, so we weren't really looking for it.

The report goes on to mention that other newspapers from the *Los Angeles News Group* began delving into the problem. The digital news director of the group confirms that some of the posts disappeared:

> On our big properties, there were missing posts, but our smaller properties didn't (have the posts disappear). We were posting (the same stories) across the board; it was kind of easy to identify the hole (…) When you're the hometown paper and you're trying to have a conversation with your readers, that's not OK (Otto, 2013).

This newspaper report coincides with the results of an interview that the author carried out in France in 2014. The Facebook page administrators (admins) of two Army sites were interviewed during a non-participant observation session. The Non Commissioned Officer in charge of one of the pages complained that several of her posts from two months ago had disappeared. She mentioned that contacting Facebook was not useful. Other page admins were asked about the same problem in the United States. One page admin from a US Army Facebook page said that many posts had been deleted by the website because "it thought it was spam." What these posts had in common

was the fact that they were missing a picture of a video and only included a link.

These interviews inspired this study. While the literature gap about this phenomenon is considerable, this study seeks to provide one of the first academic insights into the issue of fan-page-related disappearing posts. One of the tactics for this study was to contact Facebook directly. In spite of personal contact to more than one Facebook manager, it was rather disappointing to see how hermetic Facebook can be. Given this clear 'no', the alternative was to carry out an exploratory study using three different methods, three different country origins for the fan-pages and screening scholarly databases, the press, blogs and forums for further clues.

Theoretical Considerations

Fan-pages can be seen as modern instruments of institutional self-representation that code messages that are then decoded (Hall, 1980) by their audience, in this case the followers. The identity in question here is not that of the individual user that socializes with others, but that of the institution behind the fan-page. Goffman's (1959) work on individual self-representation can be projected upon an institution's image on a social media hub like Facebook. Under the idea of the 'official self' an individual is expected to hold capacities and attributes that make him or her fit for a certain situation (Riggins, 1990, p.125). The institution thus projects associations that make it suitable and even desirable for its (potential) page fans. Whether it is a brand or a public institution, the idea is to create attraction, raise awareness or meet a page-specific goal.

This attraction is created through virtual interaction

on Facebook. In order to socialize online, people normally have to represent themselves, making self-representation a condition of online participation (Thumim, 2012, p. 138). This is also relevant for institutions and not only for individual profiles, as the former also represent themselves in order to participate and generate the visibility, positioning or attraction that they seek. As Thumin (2012, p. 141) argues, a process of mediation shapes self-representation depending on the context, which is in turn influenced by "aesthetic, moral and political decisions [...] made by people other than the person representing him or herself" (p. 142). This implies that institutional self-representation is a product of a certain discourse that is projected upon the audience that decodes messages (Hall, 1980), making Facebook content a complex product of communication.

The information posted on these pages serves as messages of institutional self-representation that are triggered once the user 'likes' or 'follows' the fan page. 'Institutional power relations' (During, 1999) help shape the messages that are being broadcast or mediated (Silverstone, 2005) to an audience of followers with an underlying 'complex structure of dominance' (Hall, 1973 in During, 1999, p.90). Whether these pages are related to public institutions or company brands, they broadcast messages of attraction that build an interest-based community or a 'social circle' (Simmel, 1890).

In the interaction between organizations and their followers, the 'like' button appears as one of the first stages of virtual interaction. Users 'like' a fan page to connect their virtual profiles to its content. Following a study by the consultancy Exact Target, users ages 15 to 24 "tend to use 'Like' for purposes of self-expression and public en-

dorsement of a brand" (Exact Target, 2011, p.6). The same report states that 58% of respondents base their expectations on 'exclusive content, events or sales' (p.9), which speaks for the construction of social circles (Simmel, 1890) that help build individual identities. Thus, the way the content is projected to users, its frequency and the underlying algorithms that condition the way it is being mediated is relevant for researchers who study these virtual identities or the self-representation of institutions.

Hence, social media sites appear as hubs of interaction where meaning-making takes place constantly (Rafi, 2014). Whether we see them as further extensions of Turkle's (2005) 'second self' or as a new communication channel with underlying power structures (During, 1999), digital interaction takes place throughout the world on a daily basis, constantly constructing meaning. This digital engagement need not be entirely different from Goffman's (1959) face-to-face interactions. In his famous theater metaphor, individuals are acting just as the digital user acts before his virtual audience. Both tend to project acceptable, context-related images of themselves. We can thus visualize institutions as mediated actors 'on stage', posting content and performing their self-realization ideals while attempting to reach their organizational goals, i.e. attracting users who build their virtual identities in the process.

Methods

Notes on Data Collection

Posts can be collected using different techniques (For data crawling on Facebook, cf. Wilson et al. 2012, p. 215 and section 2 of this paper). Cormode et al. (2010) classify

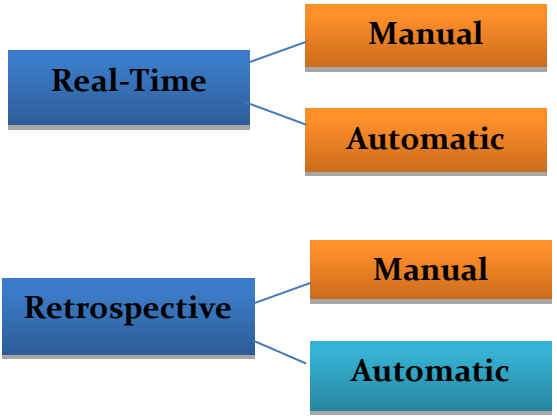

Figure 1. Data Collection Techniques

data collection in three categories (p. 5-6).

API driven: queries or calls are sent in order to fetch data on properties and relationships of the site.

Scraping based: through a web client, the researcher captures and analyzes the data extracted directly from part of the site.

Passive network measurement: through 'sniffing and parsing' (p.6), the researcher captures requests 'to and from the ONS of interest.'

In these terms, the first two scenarios explored in this paper can be seen as a way of 'scraping' data manually, while the third technique (using NVIVO 10®) collects data through a web extension, automatically scraping selected fan-pages. On the other hand, the two criteria that help differentiate the three techniques have to do with a) time (is it real-time or retrospective collection?) and b) automation (is the collection manual or automatic?). The three techniques used in this paper appear in the orange boxes in Figure 1.

Hence, the techniques can be classified as *real-time* and *retrospective* collection. Note that the study concentrates on timeline posts made by page admins, not on other fan-page content. This means that researchers can either extract data that is being produced in an approximation to real time (i.e. data from the last minutes, hours, or days) or retrospectively after more than e.g. seven days. The reliability of the data can be questioned if posts appear during real time analysis and then 'disappear' in retrospective analysis.

Study Design

Three different scenarios are explored: First, posts are extracted manually from six military fan-pages with an approximation to real-time based on a maximum of seven days. These fan pages are selected as part of a larger project involving military recruitment in France, Germany and the United States. Second, posts are manually collected after a period $t_1 > 7$ days. Third, posts are retrieved automatically every week using the tools provided by the NVIVO 10® software following $t_1 > 7$ days. The rule of seven days stems from the permanent observation of the selected six Facebook fan-pages. After seven days online, only post highlights are shown while others are hidden unless one clicks on 'all stories' for every relevant year.

The spread resulting from the three techniques is calculated in order to contrast the amount of posts collected. Let n be the numeral ascribed to each fan-page, i.e. $n = 1,2,3,4,5,6$.Each average fan-page spread is given by:

$$\delta_n = \frac{\sum_{i=1}^{T}(x_n - y_n)}{T},$$

Where i stands for the relevant days that are taken into account, ranging from 1 to T. The variables x and y stand for the number of posts retrieved in method 1 (e.g. real time) and method 2 (e.g. retrospectively).

Let σ_n be the result of weighting σ_n by the arithmetical mean of posts per day, represented by μ_n:

$$\sigma_n = \frac{\sum_{i=1}^{T}(x_n - y_n)}{T \cdot \mu_n}$$

Full overlapping of posts from the relevant collection methods is given by $\sigma_n = 0$, which implies that data retrieved e.g. retrospectively does not deviate from data retrieved e.g. in real time. The higher the spread, the less reliable it is to collect data retrospectively. Note that this comparison can also be related to automatic vs. manual real time or retrospective collection.

The distance between real time and retrospective retrieval is given by a constant $t+\propto$, whereby $\propto \geq 7$ days. Retrospective data collection was carried out after more than one month from the date of posting, in order to ensure a large distance. The same pages were screened for posts, all under the same conditions. The latter include:

- Same researcher profile (One for each country)
- Desktop Version of Facebook.
- Same Server Tunnel depending on page and country, as a means to guarantee visibility of local posts (For example, *Recrutement Armée de Terre* is only

Table 1
Basic Page-related Data

Page Name	URL	Country & Service Branch	Total Number of Likes (April, 2015)	Reach 'People talking about this' (April, 2015)
US Army Future Soldier Center	www.facebook.comArmyFutureSoldierCenter	US Army	681,002	192,313
GOARMY.COM	www.facebook.com/goarmy	US Army	909,651	37,539
Armée de Terre	www.facebook.com/armee2terre	French Army	336,893	34,076
Recrutement Armée de Terre	www.facebook.com/RecrutementArmeeDeTerre	French Army	457,294	4,435
Bundeswehr	www.facebook.com/bundeswehr	German Armed Forces	315,292	35,189
Bundeswehr-Karriere	www.facebook.com/bundeswehr.karriere	German Armed Forces	243,788	7,206

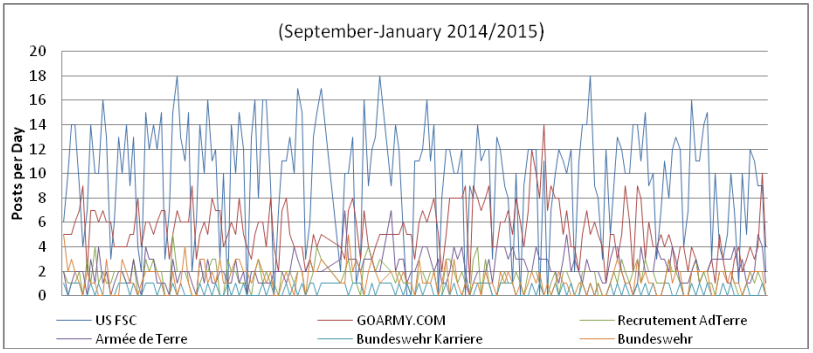

Figure 2. Real Time Data Collection

visible to 'French' users)

- Same manual screening through scroll-down.
- Filtering by 'all stories'
- Same period of time (01. September, 2014 – 23. February, 2015)

Figure 2 depicts the posting frequency as manually collected from September 2014 to February 2015. While the posting frequency of the two US fan pages is considerably higher, the four European pages rarely surpass the five posts per day mark.

Findings

Fan-Page Filters

Content displayed on Facebook fan-pages may vary depending on the chosen filters. This is especially relevant for retrospective data collection. Real-time collection avoids this problem, as fresh posts appear and do not seem to fall into the filters before the 8th day online.

Users can scroll down the fan-page filtering by *highlights* or by *all-stories*. The observations show

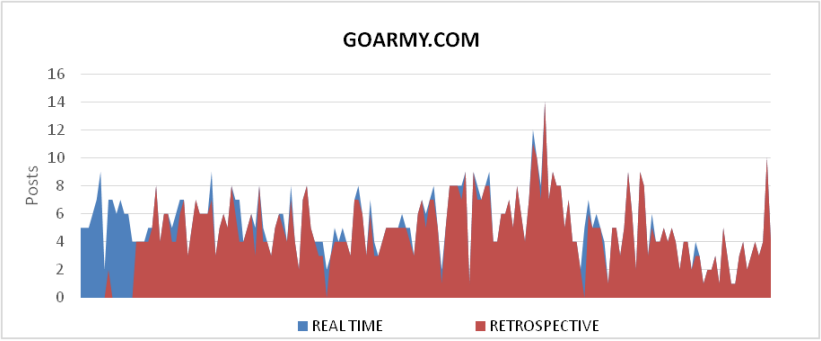

Figure 3. Filtering by 'All Stories' for 2014 and 2015

that post highlights need not be the most popular ones. The French page *Armée de Terre*, for example, displays posts with fewer than five 'likes' and not even one 'share' among the highlights. But the most important aspect of these filters is how they influence data collection. When working with two different years, say 2015 and 2014, filtering by all-stories does not apply to both years, conditioning data extraction if the second year is not filtered in the same way.

The following two figures show the different results when filtering by 'highlights' or 'all stories' before data collection. Figure 3 shows the American Page *GOARMY.COM* and the frequency of collected posts. For 2014 and 2015, posts were collected filtering by 'all stories'. The green line separates the year 2014 from 2015. In this case, real time and retrospective data seem to generally overlap (red against blue areas).

Figure 4 shows what happens if one filters by 'all stories' for one year (2015) and by 'highlights' for the other (2014). This filter reduces the amount of posts dramatically, implying that a possibly complete collection must

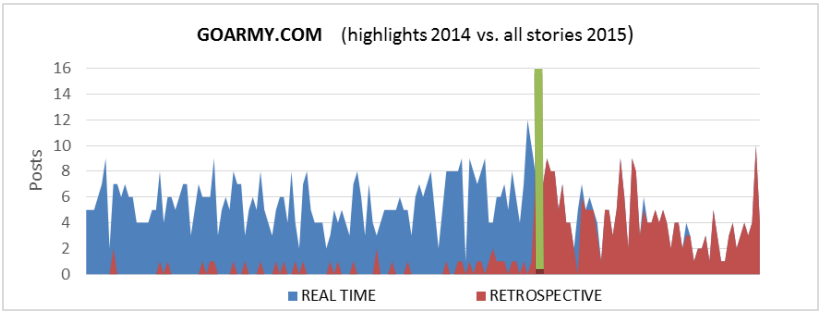

Figure 4. Filtering by 'Highlights' for 2014 and 'All Stories' for 2015

consider filtering every relevant year by 'all stories'. Figure 4 reflects a much smaller overlapping of the red and blue areas in 2014, meaning that the filter alters the amount of posts one retrieves. The observations show that this filter also impacts data collection carried out with a software program, as it can only fetch data that is not automatically hidden by Facebook. As a partial conclusion, it can be said that any collection seeking the highest possible degree of completeness must consider these filters, otherwise the researcher loses a considerable share of potentially available data. If this data is lost, conclusions derived from the available content may vary considerably.

Data Collection: results derived from three techniques

The results of the study can be classified as follows: 1) manual data collection (real time vs. retrospective); and 2) mixed data collection (real time or retrospective vs. automatic collection via NVIVO®).

Manual Data Collection (real time vs. retrospective)

The contrast between real-time manual and retrospec-

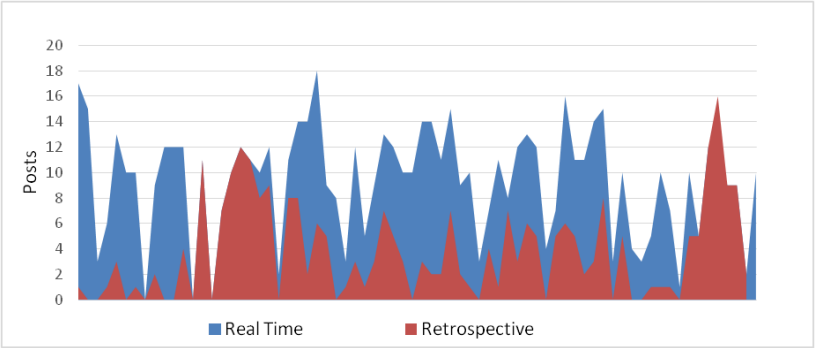

Figure 5. US page: *US Army Future Soldier Center (US FSC)*

tive manual data collection is exemplified by the following graph, which depicts the *US Army Future Soldier Center* (US FSC) page. It has the highest posting frequency in the sample. The blue area shows the posts that were collected in real time, while the red area shows the posts that were collected retrospectively, i.e. after $t+\propto$. In this case, real time and retrospective collection do not overlap homogeneously, which points to different content extraction using the two techniques. The case of the *GOARMY.COM* page (Figure 4) shows a more adjusted overlapping, which implies that the contrast of both techniques deviates across pages. This result speaks for the lack of coincidence of real time and retrospective data collection. Hence, collecting posts after several days or weeks from a fan page is not the same as collecting them in real time or shortly after the content was posted.

This contrast can be graphed using the spread given by subtracting the retrospective data from that of real time collection. The spread gives the difference in the

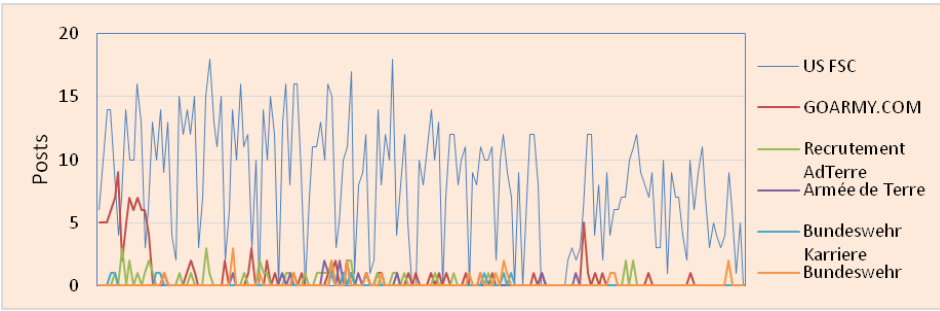

Figure 6. Real Time vs. Retrospective Spread

quantity of posts that appear by collecting posts with one method (e.g. retrospective) against another (e.g. real time). Figure 6 shows that in the case of the *US FSC* page, the spread is considerable, while the rest of the pages tend to show a spread of 0. The difference may be due to Facebook algorithms or to the individual deletion of potential spam posts. Possible explanations are subject both to further research and to Facebook's willingness to disclose information. As mentioned, according to one of the page administrators of *US FSC*, Facebook has indeed deleted many of the posts thinking that they were spam. The criterion seems to be the presence of a link, which is still usual in their postings.

In order to further analyze the spread, one may calculate δ and σ (i.e. weighted δ), as explained in the study design section. The results of the six pages are presented in Figure 7, which shows the comparably high value of the *US FSC* page as opposed to values that converge toward 0 in the five other cases. If σ is calculated by dividing by the arithmetic mean of real-time posts, i.e. how many posts are there in average, different proportions and still a comparatively high value for the *US FSC* page can be found.

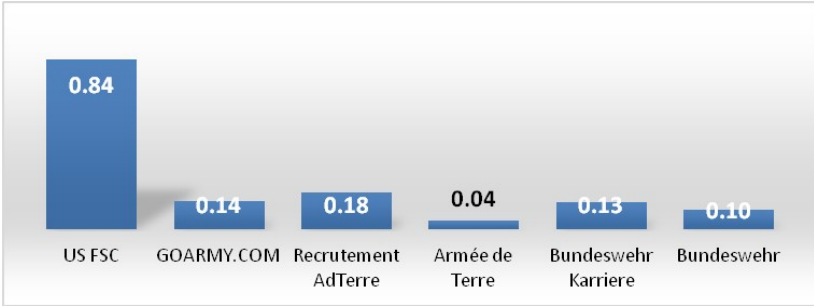

Figure 7. σ Weighted Delta of Spread (real time vs. retrospective)

The value of σ is maximized if the spread is equal to the number of real-time posts, i.e. if there are no retrospective posts.

As a partial conclusion, it can be said that real-time and retrospective collection of fan-page posts yield different results. This indicates that the researcher who is extracting data manually in an approximation to real time may view, use and base his or her research on different content than the colleague who opts to collect data from previous weeks and months. While the spread is relatively low in most of the cases seen in the sample, the page with high posting activity shows proportionally high deviations that can easily bias results. In the next section, the third technique of automatic retrieval via software is analyzed.

Mixed data collection (real time/retrospective vs. automatic via NVIVO®)

After comparing the two approaches to manual data collection, it is worth asking how the results of automatic data collection differ from those derived from manual retrieval. Since manual extraction poses transcription risks

and a great load of work for big data sets, it is sensible to compare the results of manual collection to those of a software that offers a customized tool for Facebook fan-page datasets. Using QSR's NVIVO 10® and its extraction tool NCAPTURE®, four of the six pages were scrutinized in order to extract data from the same time periods.

NVIVO® extracts datasets as .nvcx files and allows researchers to export datasets into Excel files, which can then be filtered according to specific criteria like date, user, post ID, etc. When running NCAPTURE®, the same problem of real-time vs. retrospective retrieval arises. Should one run the tool only one time (similar to retrospective collection) or several times in order to maximize the number of posts captured? During pretests, manual retrospective retrieval had already proven to be incomplete. Hence, the NCAPTURE® tool was run every week and its resulting datasets were merged by the software. This allows the technique to be classified as 'real-time'.

An example of the four different results yielded by manual real-time collection vs. automatic collection through NVIVO 10 ® can be seen in Figure 8. The results were rather surprising, as the software collection showed collection gaps in some periods for the *US FSC page*, while it almost completely overlapped in the case of the American Page *GOARMY.COM* with the exception of two gaps in September and another one in December. In the case of the German Page *Bundeswehr-Karriere,* the low posting activity almost overlaps homogeneously, but deviations are still visible. What is rather puzzling about these results is that each page shows different types of deviations. While the US FSC page shows large gaps (blue does not overlap with red), the German *Bundeswehr* (Armed

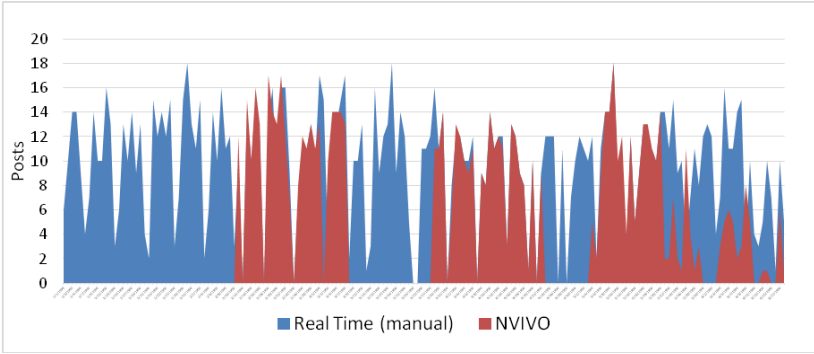

Figure 8. US Page: *US Army Future Soldier Center US FSC)*

Forces) page does not show gaps but repeated individual deviations. One possible solution to this problem, which may be addressed by further research, is the resulting dataset if the software is run on an hourly or daily basis, thus trying to maximize the number of data covered.

The differences in the four pages can be visualized by using the spread between real-time manual extraction and automatic retrieval. The following figure illustrates this spread, which is in some cases negative. A negative spread (when the line appears under the horizontal axis) means

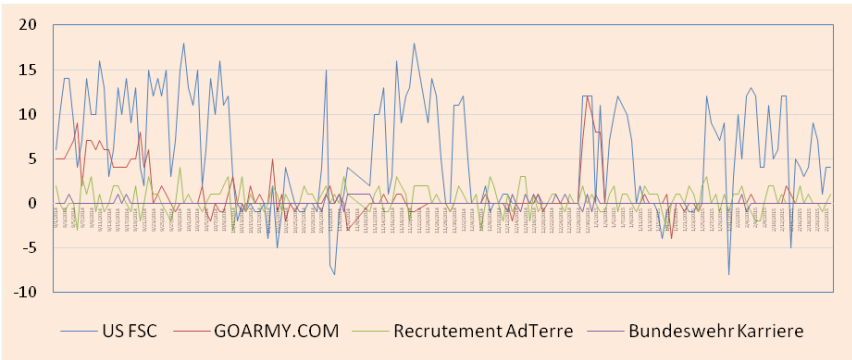

Figure 9. Real Time vs. NVIVO Spread

Figure 10. Weighted Delta of Spread (real time vs. NVIVO)

that NVIVO® has extracted, for some specific dates, more posts than the ones resulting from real-time extraction. This result can have two explanations: Either the software is able to extract content that is not visible to the researcher, or the indicated time of extraction deviates slightly between the two approaches.

The same analysis can be carried out comparing NVIVO's results with those of the manual retrospective data collection. With a more massive presence of negative spreads, we can conclude that the software extracts more posts than the researcher who manually counts the posts of the same period, pointing to a loss of data during retrospective visualization of fan-pages.

Analogous to the comparison of real-time vs. retrospective manual collection, the values σ and δ can be calculated for automatic vs. manual collection. While the spread of the *US FSC* page still surpasses that of the other pages, the difference is lower than the one portrayed above (real time vs. retrospective). Results point to a more homogenous presence for real-time extraction, while the NVIVO® results are plagued by gaps. However, these gaps may be related to the frequency in which NCAPTURE® is

run, implying that more frequent runs of the NCAPTURE® tool can yield fewer gaps. Further, the time restraint of manual extraction speaks for the use of a software.

Figure 10 shows that the weighted delta (σ) has a lower variation than the one calculated for real time vs. retrospective collection. This implies that the results of NVIVO® and those of manual real-time collection are more similar than the ones from real-time vs. retrospective manual collection. In other words, retrospective data collection carried out manually is considerably unreliable for Facebook post collection.

Conclusions and Recommendations

The pervasiveness of social media creates social dimensions that yield new possibilities for research in the social sciences. As the biggest social media site in the world, Facebook offers extensive possibilities for research. It is not only a hub of personal and institutional self-representation, but also a virtual context influenced by discourse and power relations. Its complex and massive content is of considerable value for social scientists, as it may shed light on how digital interaction and social relations acquire new dimensions.

Yet extracting data from Facebook can be rather complex and at times lead to conclusions based on biased datasets. Three data collection techniques that rely on manual and automatic scraping indicate that the content extracted from several fan-pages is different depending on how the researcher collects his or her data. While automatic collection yields more fan-page posts, manual collection can be time-consuming and subject to transcription mistakes.

These mistakes and gaps may alter results that are thought to depict how social interactions, self representation and digital meaning creation evolve. This suggests that a methodological improvement can hinder imprecise conclusions in social science.

This study suggests that real-time and retrospective data collection yield different content in terms of posting-frequency. On the other hand, the spread of the results seems to increase when the posting frequency of a fan-page augments. In other words, the more an admin posts on the timeline, the greater the difference between real-time and retrospective results. The spread is low in the case of low-frequency posting. Time filters on the timeline also alter the amount of posts considerably. If researchers do not filter by 'all stories' (even before running a software like NVIVO), they will not have access to the majority of fan-page posts. Further, exploring the six fan pages suggests that highlighted posts (i.e. posts that appear as highlights on the timeline) are not the most popular or the ones with the highest engagement.

Manual and automatic data collection also yield relatively different results. While manual real-time collection yields more posts than the automatic retrieval through NVIVO, it cannot be concluded that NVIVO is unable absorb the same data. The reason for this may be that not running the software frequently can result in gaps. The periods covered by NVIVO seem to overlap almost completely with the manual real-time posting frequency, pointing to more accuracy when frequently running the software.

These method-related conclusions shed light on the endless ways that content can be mediated by algorithms

and page-administrator decisions. If Facebook decides to carry out changes that impact what users see, their virtual interaction and perhaps even their self-representation may be conditioned, altering the virtual interactions between people, brands and institutions.

The findings allow several recommendations for social scientists who collect posts from fan-pages:

- Consider that Facebook posts are malleable symbols of self-representation and the expression of virtual interaction, which implies that they are not only dynamic but also mediated by the political and commercial context that underlies digital meaning creation. Their content is of vital importance to understand the way people and institutions socialize digitally.
- Whenever extracting posts, be sure to filter by 'all-stories' in order to maximize the amount of visible posts. This includes filtering by all-stories for *every* year in question.
- Automatic extraction via software like NVIVO or ATLAS is more efficient and avoids transcription mistakes that occur when counting and recording the data manually.
- Carry out real time, not retrospective extraction. Whether manually or automatically, real-time or an approximation to real-time yields more posts.
- If a software is being used, make sure to run it as often as possible in order to avoid gaps. Extraction via software can be best approximated to real-time extraction if the software crawls the website as often as possible. Software systems that are able to

merge datasets and replace already existing observations increase the efficiency.

Working with complex datasets from Facebook and other social media sites can yield valuable results for social scientists, yet it is essential to guarantee the reliability of the data by choosing extraction methods that maximize data completeness.

References

Bechman, A., & Vahlstrup, P. (2013). Designing Data Retrieval App to Study Facebook User Participation. *CHI'13*, April 27 – May 2. Retrieved from http://altchi.org/2013/submissions/submission_anjabechmann_0.pdf

Catanese, S., De Meo, P., Ferrara, E., Fiumara G., & Provetti, A. (2012). Extraction and Analysis of Facebook Friendship Relations in A. Abraham (Ed.), *Computational Social Networks: Mining and Visualization* (pp. 291-324). London: Springer-Verlag.

Center for Ultra-scale Computing and Information Engineering [CUCIS]. (2011). McCormick, Northwestern Engineering. Northwestern University. Retrieved from http://cucis.ece.northwestern.edu/projects/Social/

Cormode, G., Krishnamurthy, B., & Willinger, W. (2010, July 7). A Manifesto for Modeling and Measurement in Social Media. AT&T Labs-Research. First Monday, 15 (9–6), 1-17.

Crossfield, J. (2013, December 18). Beware the Social Media Content Algorithm Chasers. Retrieved April 4, 2015, from the Content Marketing Institute website: http://contentmarketinginstitute.com/2013/12/beware-social-media-content-algorithm-chasers/

Data-Extraction-Facebook [DEF]. (2009). Data Extraction Facebook: Data Mining with Facebook. Retrieved March 25,

2015 from the Google Projects website: http://code.google.com/p/data-extraction-facebook/

Digital Methods Initiative (2015). Wiki: DMI tools pertaining to Facebook. Retrieved March 26, 2015 from https://wiki.digitalmethods.net/Dmi/ToolDatabase?cat=DeviceCentric&subcat=Facebook

During, S. (Ed.) (1999). Stuart Hall: Encoding, Decoding. *The Cultural Studies Reader* (2nd Ed.) (pp. 90-103). New York, NY: Routledge.

Exact Target (2011). The Meaning of Like. *Subscribers, Fans and Followers*. Report Nr. 10. Retrieved from: http://www.exacttarget.com/resources/SFF10_highres.pdf

Facebook (2014, November 14). *An Update to News Feed: What it Means for Businesses*. Retrieved April 4, 2015 from https://www.facebook.com/business/news/update-to-facebook-news-feed

Facebook (2015). *Facebook Newsroom: Company Info*. Retrieved April 4, 2015 from http://newsroom.fb.com/company-info/

Facebook Help Forum [FHF] (2015). Why are the posts on my FB Fan Page [sic] are disappearing but appear on my Timeline. Retrieved April 4, 2015 from https://www.facebook.com/help/community/question/?id=10200665928150762

Gjoka, M., Kurant, M., Butts, C., & Markopoulou, A. (2010). Walking in Facebook: a case study of unbiased sampling of OSNs. *IEEE Infocom*, 1-9.

Gjoka, M., Sirivianos, M., Markopoulou, A., &Yang, X. (2008). Poking facebook: Characterization of OSN applications. *Workshop on Online Social Networks (WOSN'08),* 31-36. Retrieved from http://conferences.sigcomm.org/sigcomm/2008/workshops/wosn/papers/p31.pdf

Goffman, E. (1959). *The Presentation of Self in Everyday Life* (1st Ed.). New York, NY: Anchor Books.

Hall, S. (1980). Encoding/Decoding. In S. Hall, D. Hobson, A.

Lowe & P. Willis (Eds.), *Culture, media, language: Working papers in cultural studies, 1972-79* (pp. 128-138). New York, NY: Routledge.

Manning, C., Raghavan, P., & Schütze, H. (2009). An Introduction to Information Retrieval, Online Edition [Draft]. Cambridge: Cambridge University Press. Retrieved from http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Milolidakis, G., Akoumianakis, D., Kimble, C., & Karadimitriou, N. (2014). Excavating Business Intelligence from Social Media. In J. Wang (Ed.), *Encyclopedia of Business Analytics and Optimization*. (pp. 897-908). Hershey, PA: IGI Global.

Nasution, M. K. M., & Noah, S. A. (2012, July 16). Information Retrieval Model: A Social Network. Extraction Perspective. Retrieved from: http://arxiv.org/pdf/1207.3583.pdf

Otto, F. (2013, August 10). Disappearing Facebook posts puzzle and concern editors. *The Mercur*y. Retrieved from http://www.pottsmerc.com/article/MP/20130810/NEWS01/130819954

Rafi, M.S. (2014, May 14). Meaning Making Through Minimal Linguistic Forms in Computer-Mediated Communication. *SAGE Open*. 4 (201X), 1 –12. http://dx.doi.org/10.1177/2158244014535939

Rieder, B. (2013). Studying Facebook via data extraction: The Netvizz application. *WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference*, 346-355, New York, NY: ACM.

Riggins, S. H (Ed.) (1990). Beyond Goffman: Studies on Communication, Institution, and Social Interaction. *Approaches to Semiotics Series,* 96. New York, NY: De Gruyter.

Silverstone, R. (2005). The sociology of mediation and communication. In C. J. Calhoun, C. Rojek & B. S. Turner (Eds.), *The SAGE handbook of sociology* (pp. 188-207). London:

SAGE Publications.

Simmel, G. (1890). *Über sociale Differenzierung: sociologische und psychologische Untersuchungen* (1st Ed). Leipzig: Duncker & Humblot.

Thumim, N. (2012). *Self Representation and Digital Culture*. New York: Palgrave Macmillan.

Traud, A. L., Kelsic, E.D., Mucha, P.J., & Porter, M.A. (2009). Community structure in online collegiate social networks. *Society for Industrial and Applied Mathematics Review*, 53 (3), 526–543. http://dx.doi.org/10.1137/080734315

Turkle, S. (2005). *The Second Self: Computers and the Human Spirit* (Twentieth Anniversary Ed.). Cambridge, MA: MIT Press.

Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. (2009). On the evolution of user interaction in Facebook. *Workshop on Online Social Networks (WOSN'09)*. Retrieved from: http://www.mpi-sws.org/~gummadi/papers/wosn23-viswanath.pdf

Wilson, R.E., Gosling, S.D., & Graham, L.T. (2012). A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science*. 7 (3), 203-219.