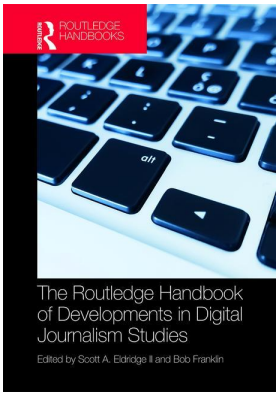
This article was downloaded by: *10.3.98.93*
On: 17 Jan 2019
Access details: *subscription number*
Publisher: *Routledge*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: 5 Howick Place, London SW1P 1WG, UK



The Routledge Handbook of Developments in Digital Journalism Studies

Scott A. Eldridge, Bob Franklin

Reconstructing the Dynamics of the Digital News Ecosystem

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315270449-10
Elisabeth Günther, Florian Buhl, Thorsten Quandt
Published online on: 30 Aug 2018

How to cite :- Elisabeth Günther, Florian Buhl, Thorsten Quandt. 30 Aug 2018, *Reconstructing the Dynamics of the Digital News Ecosystem from:* The Routledge Handbook of Developments in Digital Journalism Studies Routledge
Accessed on: 17 Jan 2019
https://www.routledgehandbooks.com/doi/10.4324/9781315270449-10

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

9

RECONSTRUCTING THE DYNAMICS OF THE DIGITAL NEWS ECOSYSTEM

A case study on news diffusion processes

Elisabeth Günther, Florian Buhl, and Thorsten Quandt

Introduction: preconditions to observe news diffusion processes in the digital news ecosystem

News diffusion processes as the conceptual frame at the ecosystem level

The digitization of news production, distribution, and consumption has provided journalism researchers with plenty of opportunity to explore the establishment of new structures and processes within the developing networked media system (Franklin and Eldridge, 2017). Alongside rather manifest distinctions of online journalism, such as the reorganization of newsrooms, journalism scholars have started to discover digital news phenomena whose accessibility requires the advance of research designs in the first place. Observers might recognize these phenomena without the help of innovative research methods. Readers might recall personal experiences with online journalists' publish-first-and-update-later routines (Karlsson and Strömbäck, 2010; Saltzis, 2012; Widholm, 2016) or the acceleration of the news cycle (Rosenberg and Feldman, 2008). However, transferring such anecdotal evidence into a set of more systematic, generalizable observations of such phenomena is a reconstructive endeavor that benefits from the interplay of methodological advances and theoretically focused procedures of data analysis.

In this vein, Zamith (this volume, Chapter 7) presents innovations in content analysis that enable journalism researchers to systematically trace the updates of online news items and websites (see also Karlsson and Strömbäck, 2010; Saltzis, 2012; Widholm, 2016). While we share the shift from a static to a dynamic perspective, our main focus is on the digital news ecosystem as a whole – or subsections, at least – and on discovering the potential dynamics of news diffusion processes among the variety of providers within. Our approach is inspired by the idea that online journalists' reaction times to emerging stories is getting shorter and shorter in the wake of vanishing publication deadlines (Risley, 2000) – due to technological opportunities of live reporting (see Artwick, this volume, Chapter 22), the normalization of immediate coverage beyond major news events (Lim, 2012; Rosenberg and Feldman, 2008), or the orientation toward the latest stories of competing news sites (Boczkowski, 2010). As the trend toward immediate event coverage appears to be a global attribute of digital newswork, we aim to complement studies observing

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

the routines in single newsrooms (Boczkowski, 2010) with a research design that reconstructs the flow of information in the digital news ecosystem (Weber and Monge, 2011) and especially its dynamics. If a multitude of online newsrooms tend to cover newsworthy events instantaneously, the routines of individual newsrooms will lead to close timing of event-related publication decisions at the ecosystem level. Conceptually borrowing from the literature on the diffusion of news in the public (Rogers, 2000), we developed a research design to capture resulting news diffusion processes in the digital news ecosystem (Buhl et al., 2018). Within this analytical framework, we can map the dynamics of these processes, as we relate the amount of time elapsed since the first report on an event with the subsequently accumulating number of news sites covering the story, too.

We aimed first to establish and systematically capture the phenomenon of digital news diffusion processes and second to analyze the conditional factors of their dynamics (see Buhl et al., 2016, 2018).

Computational methods enable the reconstruction of news diffusion processes

Similar to tracking the velocity of online news items (see Zamith, this volume, Chapter 22; see also Karlsson and Strömbäck, 2010; Widholm, 2016), capturing the dynamics of news diffusion processes among online news sites relies heavily on recent advances in automated content analysis. News diffusion processes can be uncovered only when shared routines of prompt news dissemination in digital journalism converge with innovative methods of content analysis in journalism research, which are both highly sensitive to time and scalable to the ecosystem level of analysis. Real time data collection of news items published by the variety of providers within the same digital news ecosystem is a precondition for the reconstruction of diffusion processes. We believe it is fair to say this analytical prerequisite would push manual coding to its logistic limits (see Figure 9.1). Additionally, the project's 'catch all – select later' strategy requires data storage in a searchable database, because researchers can identify relevant events only *ex post facto*.

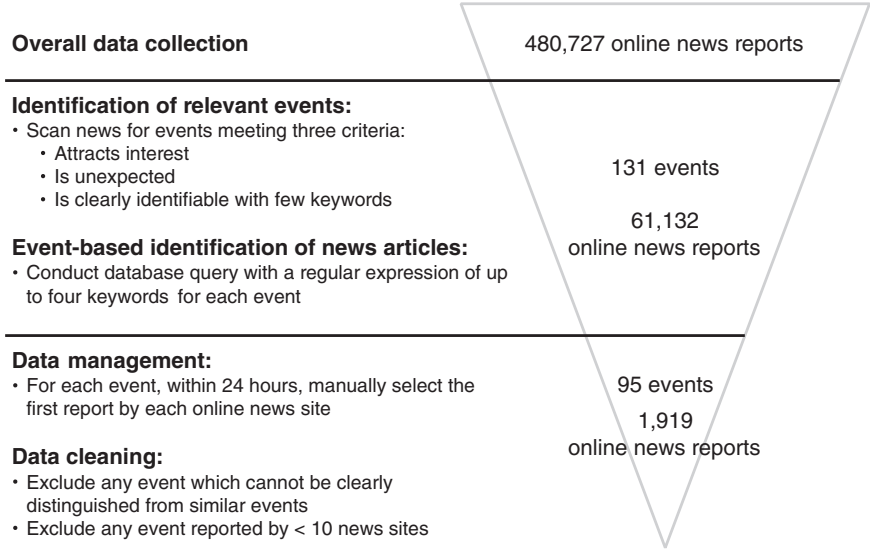


Figure 9.1 Stepwise procedure to identify relevant events and related online news reports (Buhl et al., 2018)

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

We compiled a research design meeting these requirements in our work on the dynamics of diffusion processes among professional online news sites in Germany (Buhl et al., 2016, 2018), which we will use as an example throughout this chapter. However, we will generalize from this case study to provide readers with the basic principles of systematically capturing diffusion processes within the digital news ecosystem. In the subsequent section, we will detail the requirements for data collection and provide readers with solutions for each step during the process. In the third section, we will turn to the data analysis, first outlining how to reconstruct digital news diffusion processes and how to systematically describe their dynamics. Thereafter, we will introduce analytical procedures both to identify subgroups of diffusion processes that share similar dynamic patterns and to test potential relationships between event attributes and diffusion dynamics. In the final section, we will discuss links of the news diffusion process-approach to similar concepts as well as potential future advances of studying the dynamics of digital journalism from an ecosystem perspective.

Tracking news diffusion processes online: basic research design

Over the past decades, both object of study and methods of analysis in digital journalism research have been transformed by the digital revolution. Tracking news diffusion processes online is a good example for the changes that need to be implemented on both sides of the research process – concerning both *what* and *how* – and a challenging task at multiple levels. In the following, we first outline challenges and possible solutions related to the blurry boundaries and fluidity of content online (What is the digital news ecosystem? How do we perform data collection on relevant content?) and then turn to the specific unit of analysis (What are the markers by which to define relevant events? How can online news diffusion processes on an event be reconstructed for analysis?).

Sampling unit: representing the online news ecosystem

Defining the boundaries of the online ecosystem is a Sisyphean task: online, legacy media are embedded in a complex network of information providers, with multiple ties to and from social media (Leskovec et al., 2009), internet-specific news formats (e.g., BuzzFeed, or news aggregators such as Reddit), and of course traditional sources and distribution channels offline. Links are not always made explicit and thus further obscure the shape and boundaries of the news ecosystem. Consequently, the list of relevant websites for a study on news diffusion online is, to a certain degree, open to discussion. Combined with the sheer size of available information online, it is at this step inevitable to make a compromise of some sort. In terms of defining the sampling unit, we identify two main strategies.

Researchers interested in *the interplay between different types* of news providers might prioritize diversity over completeness. By drawing a sample of a few sources from many different formats, this allows for anecdotal insights into their role and strategy. Researchers who aim to explore a specific phenomenon, as we do in our case study, might follow the opposite strategy and focus on reproducing a *comprehensive model of one domain* only. By targeting an internally consistent fraction of the information cosmos (in our case, German print newspapers' online outlets) we can benefit from two main advantages. First, this approach provides a common baseline for assessing in-group differences and identifying characteristic subgroups. Second, basing the sample on a predefined number of websites makes it feasible to add meta-information on each of the providers, enabling us to control for variations in their professional background and working routines. When working with a high variety of sources, as accomplished by the first approach, the workload associated with this step is likely to get out of hand and thus push the logistic boundaries of a project (Waldherr et al., 2017).

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

For our case study, the second strategy provides an adequate starting point to explore the ways traditional news providers have adapted to the challenges presented by the 24-hour news cycle. Instead of taking on the impossible task of trying to cover all news websites online, we start with the well-structured set of print newspapers and work our way forward from there: key data is provided by Schütz (2012), listing all print publications for the German market as of 2012. For each entry, we investigated whether an online news site with an RSS (Real Simple Syndication) feed was provided. This feature provides considerable benefits with regard to a reliable data collection (as explained in the following section) and, since RSS is a common online standard, did not noticeably compromise our sample. This way, we constructed a sample of 28 websites, covering all major news outlets in Germany (for a detailed list, see Buhl et al., 2018).

Data collection: considerations

Now that we are equipped with a list of news websites to represent the online news ecosystem, the next challenge is to develop a plan for the data collection. Given the fluidity of online content, it is crucial to consider possible complications ahead of time and test the practicability of one's approach in a pretest. We follow a catch-all strategy, meaning that we collect all news articles published by the given websites over the course of a given time period *before* selecting the actual units of analysis, i.e., the events for which we then reconstruct digital news diffusion processes. There are two main reasons behind this decision. First, we are interested in *unexpected events* (our rationale for this criterion is explained in the following section), so we cannot predict their beginning and/or end.

Second, pretests indicate that online news diffusion processes cannot reliably be recovered *ex post facto* – even if only a few days or even just several hours had passed. This is mainly due to online journalists' routine of updating, rewriting, or deleting content during the initial time span in which news on unexpected events unfolds, adapting their stories to new information in real time (Karlsson and Strömbäck, 2010; Saltzis, 2012; Widholm, 2016). If not made transparent, such live editing can leave researchers confused: Looking at published articles in hindsight, it might appear as if one news outlet was hours ahead of others, with detailed coverage on an event at a point in time when others only had a note along the lines of "something has happened, we will elaborate shortly". In the highly competitive environment that the online news ecosystem constitutes, such a scenario is unlikely. Even in a case where one newsroom is given a head start due to exclusive access to information, other websites will catch on quickly; online, co-orientation among news providers has become both easier and, with immediacy being a crucial production norm, more important (Boczkowski, 2010; Karlsson and Strömbäck, 2010). A more plausible explanation is that the timestamp of the initial post has remained unchanged throughout later edits. If the research interest is to reconstruct the first substantial report on a specific event for each news website, as we plan to do, this situation makes continuous and (near) real-time data collection necessary. For our case study, we therefore set up a comprehensive data collection over the course of nine months (June 2013 to March 2014), making sure that seasonal effects, such as variations in working routines over the holiday season in December or the so-called 'silly season' over the summer months, can be ruled out, and that a sufficient number of relevant events are covered within the sample period.

Data collection: technical setup

Both the reliable collection and the efficient storage of online content are associated with certain technical challenges. While web-scraping is surely not a trivial task, there is good introductory material available to guide motivated non-IT researchers through the basic steps of this process.

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

Trilling (2017) provides a good example of a step-by-step guide, handholding readers through the lines of a custom Python script. Collecting news articles from a couple of websites over the course of a week is, with such introduction and a little practice, doable even for beginners. Widening the scope both regarding the number of sources to be monitored simultaneously and regarding the time frame of the data collection (over several weeks, months, or even years) will, however, quickly lead to a data volume that requires a more advanced setup.

The task of reconstructing online news diffusion processes, as presented in our case study, can consequently be characterized as a *Big Data project*. Setting up our data collection, several technical aspects had to be considered: (1) As described, we need to download new articles regularly, aiming for at least an hourly interval, to ensure that quick developments are captured in near real time. (2) One round of downloads, preferably including preprocessing and data cleaning, has to be completed within this time frame to avoid "collusion" with the following rounds, otherwise risking a delay. Due to the aforementioned journalistic routine of re-editing published content, regular and continuous data collection is essential to guarantee the reliability of our project. (3) Given the number of steps involved in each round (access RSS feeds, identify links to news articles, extract article metadata, download raw HTML content, save to database, perform data cleaning and preprocessing, etc.) and the repetitive nature of the overall process, performance needs to be optimized to the level of split seconds per task, since they quickly add up to serious delays. (4) In times of breaking news coverage, the article volume per hour is expected to go up. On top of all previous considerations, this means the whole process needs to be implemented in a way that does not already exhaust the available resources (time and processing power) during standard operation. This way, we make sure that the data collection stays in running order and does not slow – or even worse – break down during the high-intensity phases that are of special interest to us.

The key concept here is *scalability*: for example, saving 1 second per article does not sound like much, but means that we can save 5 minutes for the simultaneous download of 300 articles – a plausible number for an hourly collection of 30 news websites. Double the number of articles (which might happen during especially busy morning hours), and we save 10 minutes for just one task in our data collection pipeline. One of the main technical levers to optimize performance lies in parallel processing, meaning that independent tasks are efficiently distributed across all available CPU cores of the system. Applying parallel processing to each step can thus quickly cut working time to a fraction: before optimizing performance, our data collection (without preprocessing and cleaning) took 30 minutes – a task that is now[1] completed in less than 2 minutes.

Depending on the specifics of the project (number of sources, time frame for data collection, hardware specs, skills, and experience of the researchers involved), arising challenges might move the task of the automated data collection from a social science pet project into the field of expertise of an IT professional. In our case, we rely on the *datamessie* open-source project (available on GitHub under the GNU General Public License). In cases where a custom solution is required, collaborations with technically skilled colleagues or even professional contracting is recommended for beginners. For advice on database selection and setup, Günther, Trilling, and van de Velde (2018) introduce relevant options and considerations. In any way, we would like to stress the point that given the fluidity of online content, a regular and timely monitoring of the data collection is key to avoid irreversible loss of information.

Unit of analysis: selecting 'relevant' events

Based on the described setup, our continuous hourly data collection of 28 news websites over the course of nine months results in an overall sample of $N = 480{,}727$ news articles. Not all are relevant to our analysis. The next step is to identify relevant events that occurred within the

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

observed time frame and which we will then trace in our sample to reconstruct their news diffusion processes. The 'relevance' of an event is not easy to operationalize and is contingent on the specific research question – for our case study, we defined three criteria to guide this process: (1) The event is newsworthy enough to trigger news coverage from a sufficient number of websites in our sample. This criterion establishes that there is broad enough interest within the online news ecosystem, allowing us to observe and compare varying reaction times across the websites in our sample. (2) The event is unexpected, meaning that news coverage has a clear but unknown starting point. As an example, the Winter Olympics 2014 are certainly a remarkable event but have been announced and discussed months ahead of the actual opening ceremony. This makes it difficult to isolate reports on the event itself in the following analysis phase. In contrast, an underdog winning a gold medal in one of the competitions might trigger the wide and easily traceable news diffusion process we aim at. (3) The event can be described with a few distinctive keywords. This is a technical prerequisite: Given the size of our sample, we need to keep the manual workload in the following selection phase to a limit and will therefore filter reports on each event by means of an automated keyword search.

By evaluating every downloaded news article from two online news sites under study against all three criteria, we identified 131 possibly relevant events within the sampled time frame.

Putting it all together: reconstructing news diffusion processes

With all necessary information at hand, we can start processing and analyzing the collected data (see Figure 9.1). First, we are equipped with a large collection of all news articles published by the observed 28 websites in the sample period, stored in a database, and second, we have a tailored list of 131 potentially relevant events that occurred within this time frame. The analysis builds upon both sets of information by retrieving all news articles that make a reference to one of the selected events, technically implemented by means of an automated keyword search. This requires basic database skills, such as a working knowledge of SQL and a fundamental understanding of regular expressions (Friedl, 2006). For each event, we defined a regular expression with one or two mandatory keywords and up to two optional keywords: to be identified as event related, an article must mention both keyword1 and, if specified, keyword2, in combination with either keyword3 or, if specified, keyword4 (see Appendix A in Buhl et al., 2018 for the final list of keywords for all events). While this sounds rather technical, it is not complicated when put into action: For example, the birth of 'Royal Baby' Prince George is described by a combination of 'William' with either 'birth' or 'son' – any news article that matches these criteria is filtered for further analysis. Keywords typically referred to the *who* – including named entities for the main actors involved in an event (names of persons, institutions, and organizations such as the Italian Senate), the *where* – names of the place or scene (e.g. London Heathrow airport), and the *what* – dense descriptions of the event if available (e.g., birth, train accident). Since our interest is constrained to highly newsworthy and unexpected events, finding clear keywords for them was mostly a straightforward process. To avoid overlooking important cues, we systematically double-checked and refined all keywords in a manual pretest. Although manual reevaluation is costly, we recommend erring on the side of caution and choosing keywords that rather yield too many false positives than to miss important articles with different phrasing. In addition, certain information that perfectly describes an event in hindsight might not have been available yet in the initial stages of reporting.

Finally, we joined the selected keywords to a regular expression and then conducted a database query for matching news articles (see Figure 9.1). In our case study, this query resulted in 61,132 matches for 131 events – a considerable reduction, but still a lot of information to be processed. For each event, we now manually selected the very first news article by each news site (within a 24-hour limit after the first report was published), which also yielded information on the overall

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

number of websites that reported on the event in question. Given our criteria for relevance as described in the previous section, aimed to restrain the scope to events characterized by high newsworthiness, we determine a minimum range of the diffusion rate and removed all events that were picked up by less than 10 news outlets. This way, we constructed a final sample of 1,919 news articles regarding 95 events. As with the data collection itself, this massive reduction of information brings with it a high workload for pretesting and double-checking to ensure a reliable research process.

Analyzing the dynamics of news diffusion processes

While the concept of news flows might seem straightforward at first, there are different ways to look at it (we will expand on this aspect later). In our case study, we conceptualized news diffusion processes as the relationship between the amount of time elapsed since the particular event has been reported for the first time (*x*-axis) and the accumulated number of online news providers in the same news ecosystem having reported the event so far (*y*-axis, see Figure 9.2). The diffusion of stories among news providers is calculated by continually adding news sites at the time lags of their first reports on the specific story. As we were able to determine publication times to the split second in our case study (Buhl et al., 2018), usually there was only one news site joining the diffusion process per time frame. But we also found instances of several news sites releasing their stories at the very same point of time (we assume because some local news sites share national newsdesks), so that the accumulation grew by multiple newspapers in that particular second.

Literally, you can best get a picture of the dynamics of digital news diffusion processes by plotting them. Figure 9.2 displays the curves of diffusion processes of three exemplary events from our case study on professional online news sites in Germany (cf. Buhl et al., 2018): the birth of "Royal Baby" Prince George (circles), the release of allegations of child abuse against Woody Allen (triangles), and a Texas state court decision declaring a ban on same-sex marriage as unconstitutional (squares). At first glance, the three cases do not seem to be that different: Within the 24-hour time frame observed for each diffusion process, it takes 13.7 hours for the last online news site to report the birth of Prince George, 9.8 hours in the case of the allegations against Woody Allen, and 14.0 hours in case of the ruling on same-sex marriage in Texas. Comparing the range of the diffusion processes, i.e., the total number of online news sites having reported each event after 24 hours, we find a slightly stronger hint at differentiated patterns of diffusion processes. There were 26 online news sites reporting the birth of the 'Royal Baby', 21 sites covering the Woody Allen story, and only 11 sites reporting the Texas court decision.

From the plots of the three diffusion curves (see Figure 9.2), however, it appears obvious that specifying *length* and *range* of complete digital news diffusion processes does not tell the full story about their dynamics. Contrary to the process patterns of the diffusion of news about the allegations against Woody Allen and the Texas court ruling, the diffusion curve for the 'Royal Baby' approaches saturation many hours prior to the final completion of the process. That is, during diffusion processes of this type, the vast majority of online news sites issue their first reports on the corresponding event in the direct aftermath of the incident – or its very first coverage in the digital news ecosystem, respectively. During this short time frame, the number of news sites joining the diffusion process grows at very fast *pace*, while it slows down significantly for the remaining time span. We found this dynamic to be a common pattern in our case study on diffusion processes within the ecosystem of German online news sites (cf. Buhl et al., 2018). For the curves of 68 from a total of 95 diffusion processes, we could identify a shift from an early burst to slower rates of accumulation. To describe news diffusion processes accurately, we consequently need to differentiate the *main diffusion phase* from the process as a whole. To determine

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.

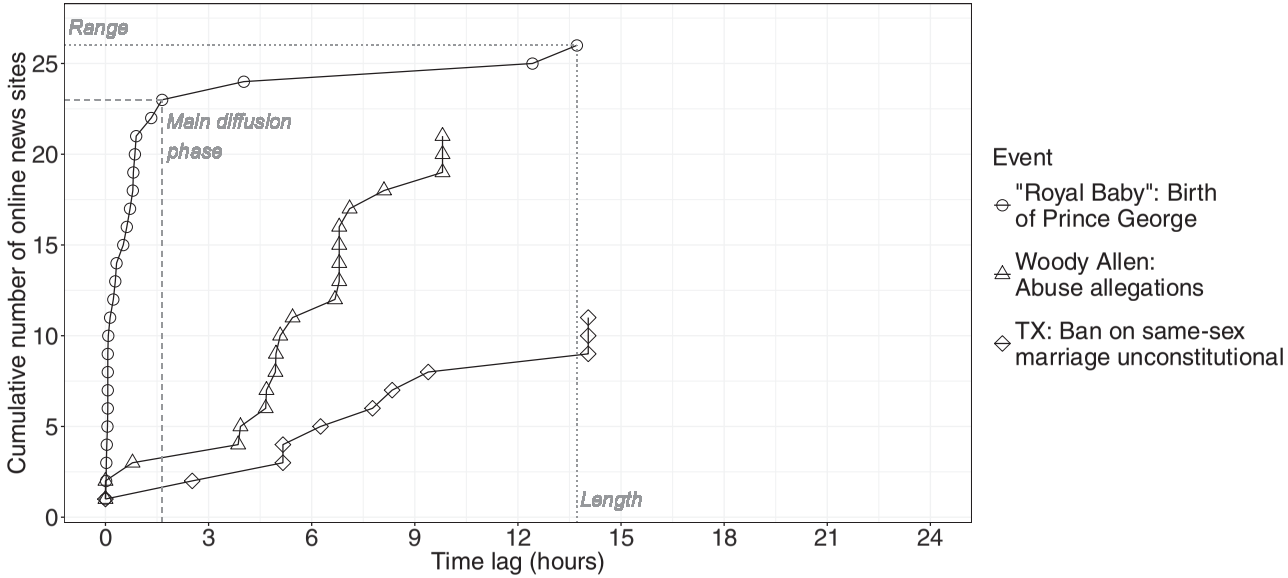


Figure 9.2 Curves of the digital diffusion processes of three exemplary events

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

the point of shifting accumulation pace for each curve, we had to rely on visual inspection in our own study, because the number of data points per curve – the total number of news providers reporting each event – was too small for the application of more sophisticated procedures. For the 'Royal Baby' story, for example, we found the endpoint of the main diffusion phase to be reached after 1.6 hours, including 23 online news sites (see Figure 9.2). For processes lacking a clear transition from a surge at the beginning to slower accumulation rates, e.g., the diffusion of the Woody Allen story and the Texas court ruling, one option is to define the main diffusion to last for the full process (as we did), so that the endpoint of the complete process coincides with the endpoint of the main diffusion phase.

Exploring dynamics of online news diffusion processes in large data collections

The groundwork regarding data preprocessing and cleaning, combined with the set of computational methods that have been outlined in the previous section, now allows us to apply this procedure to many more diffusion processes from the same digital news ecosystem. Therefore, we can widen our analytical perspective beyond describing single case studies of digital news diffusion processes in their own right to mapping their general temporal patterns across a variety of cases (see Figure 9.3 for the sample of 95 digital diffusion processes from our own research). From this global point of view, we consider each reconstructed diffusion process as a case and the whole of them as the sample. Once we have extracted information about length, range, and duration of the main diffusion phase from each individual process, we can provide frequency tables and histograms for each measure and calculate their central tendency and variability for the whole sample of processes – or for subsamples, if we wish. In our own research, we made all time-related calculations exact to the second, the time format of our time-lag data, but we also reported the results converted into hours for better readability (Buhl et al., 2018).

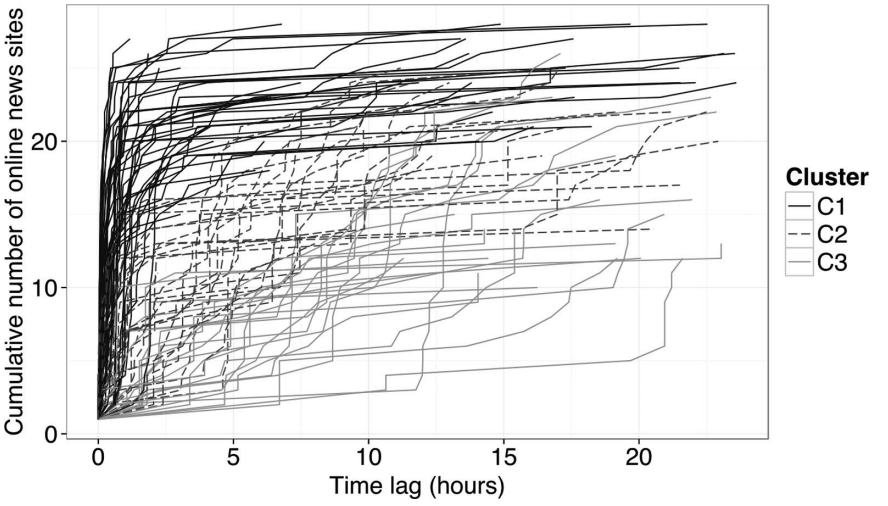


Figure 9.3 Diffusion curves for 95 events as reported by German online news sites, grouped by dynamics-based cluster membership (Buhl et al., 2018)

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

In our research of diffusion processes among German online news sites, we found the distribution of lengths of main diffusion phases to be strongly skewed to the right, with its mean at 5.1 hours (18,219 seconds) and its median at 2.2 hours (7,916 seconds), confirming the commonness of bursts and surges at the beginning of diffusion curves. To make the specific diffusion dynamics accessible beyond the endpoints of main diffusion phases, we statistically reconstructed the average diffusion curves, which was done by pooling the time-lag data from each diffusion process belonging to the same subset into a virtual, aggregated diffusion process. The pooled data can be rearranged from small to large time lags with the help of procedures of event history analysis, e.g., according to the Kaplan-Meier estimation method. The resulting aggregated diffusion curves represent the typical diffusion-process dynamics in the sample. They provide the proportion of online news sites having reported events at specific time lags, visualizing the fast accumulation of reporting sites at the beginning of diffusion processes, for example. This analytical perspective becomes especially informative as we start comparing the dynamics among subsamples of diffusion processes (see examples in subsequent sections).

Structuring samples of digital news diffusion processes by diffusion dynamics

One of our research questions was aimed at identifying recurring and common dynamics of digital news diffusion processes. The three exemplary diffusion processes plotted in Figure 9.2 indicate that these dynamics are not restricted to a single pattern. To account for this variety of dynamics, we first employed an inductive procedure based on process dynamics. We divided the full sample of diffusion processes into subsamples featuring similar dynamics by clustering the diffusion processes. The resulting process clusters serve as an indicator of common, recurring diffusion dynamics in the digital news ecosystem.

In our research on diffusion processes among German online news sites (Buhl et al., 2018), we conducted a hierarchical agglomerative cluster analysis. We treated each diffusion process as a case and used the full sets of time-lag data from each process. We deliberately decided against controlling for the length of the processes, as we consider this information a crucial differentiating factor among diffusion dynamics. The results of the cluster analysis suggests a three-cluster solution, i.e., a differentiation among three types of common diffusion dynamics (see Figure 9.3): high-range diffusion processes characterized by bursts briefly after the very first reports about the respective incident (Cluster 1, $n = 43$, e.g., the birth of Prince George); processes featuring less distinct bursts at their beginning and lower range both in total and relative to the amount of time elapsed (Cluster 2, $n = 28$, e.g., the publication of abuse allegations against Woody Allen); and processes accumulating slowly during longer time frames (Cluster 3, $n = 24$, e.g., the Texas state court ruling in favor of same-sex marriage). To characterize the process dynamics of each homogeneous subsample and to make comparisons among them, summary statistics of range, length, and duration of main diffusion phases for each subsample provide valuable insights. Additionally, we can reconstruct typical diffusion dynamics in each subsample with the help of the data-pooling procedure outlined in the previous section. For each of the three dynamics-based clusters of diffusion processes we found in our study, Figure 9.4 displays the curves of the resulting aggregated diffusion processes. The display of differentiated diffusion dynamics of subsamples offers a good starting point for discovering *post hoc* explanations for each pattern. Beyond high newsworthiness, we found events that follow the dynamics of Cluster 1 normally do not require journalists to add contextual information to be easily understood by the audience, for example.

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

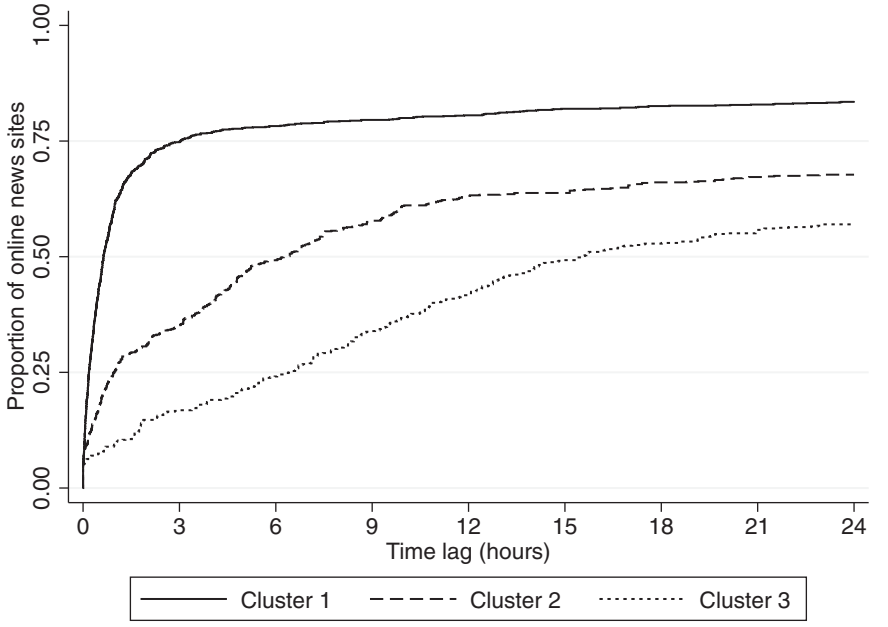


Figure 9.4 Aggregated curves for three clusters of digital diffusion processes featuring similar dynamics, from a total of 95 diffusion processes among German online news sites (Buhl et al., 2018)

Comparing the dynamics of online news diffusion processes by process attributes

If prior knowledge about the logics within the digital news ecosystem allows researchers to outline hypotheses about factors influencing the dynamics of diffusion processes, this information needs to be collected systematically. Factors may be attributes of events (either necessitating or inhibiting immediate coverage) or conditions of working routines in online newsrooms varying between diffusion processes. To explain diffusion dynamics in our own work on German online news sites (Buhl et al., 2016), we analyzed relationships with the news factors of events as well as with starting points during daytime vs. during the night, when online newsrooms are typically less staffed. To test the assumed variations, we again relied on event history analysis according to the Kaplan-Meier estimation method. Aggregated diffusion curves were reconstructed for subsamples of diffusion processes sharing the same predefined attribute. As a result, we can describe the typical diffusion patterns of processes sharing the same attribute and compare subsets of processes with different attributes. Figure 9.5 displays the result of this analysis for the time of day a diffusion process set off as a contingent condition of digital newswork. Diffusion processes starting between 10 a.m. and 10 p.m. (Berlin time) were more likely to exhibit distinct bursts at their beginning, on average having already involved 12.5 news sites after just one hour. For processes starting during the night (10 p.m.–10 a.m.), the diffusion rate is slower during the first couple of hours (7.2 news sites after one hour). As visible in Figure 9.5, they need more time to catch up to the range of day processes.

Conclusion and directions for future research

Computational methods allow us to explore dynamic environments, such as online news websites, and to take a snapshot of the ever-changing online news ecosystem. In this chapter, we

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

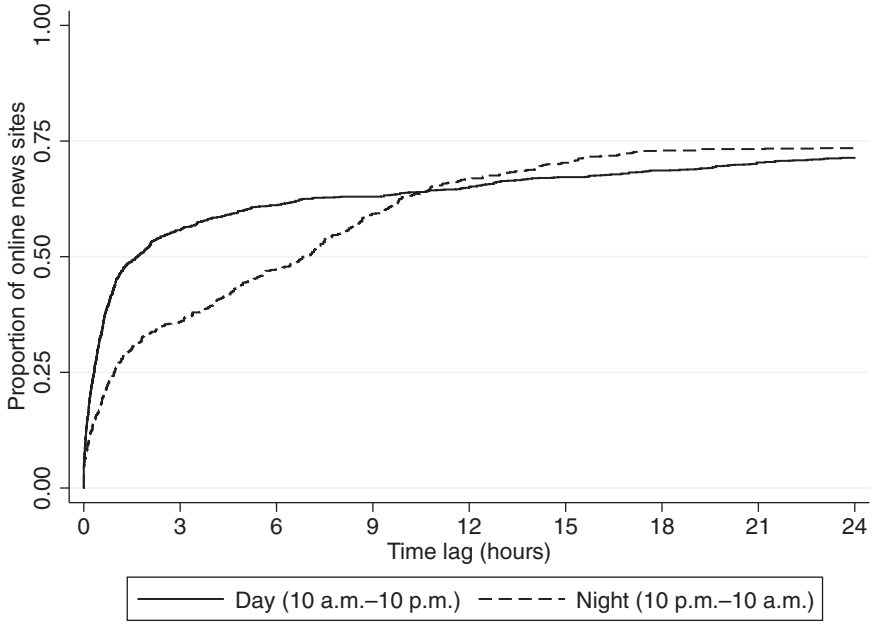


Figure 9.5 Dynamics of aggregated diffusion processes, grouped by the time of the day they were first reported (Buhl et al., 2016)

presented our take on this complex task, as illustrated by a case study on diffusion processes among German online news sites.

It is important to point out that the processes reconstructed in our case study might not necessarily qualify as *news flows*, however tempting this assumption might be. Observing time lags among various online news outlets reporting the same story, as we do, may indicate a co-orientation among journalists in various newsrooms, resulting in the circulation of a story. But time lags might also be caused by independent, parallel processing of information from the very same source. Time lags then indicate varying speed, for example due to the availability of staffers during different times of the day or variations in strategy and fact-checking routines (see also Harder et al., 2017). Especially when focusing on the coverage of major news events, as we do, we expect little uncertainty among journalists about whether or not to cover the story, so that they are unlikely to back up their news decisions by co-orientation processes (Donsbach, 2004).

There are many other possible – and equally valid – ways to approach and analyze dynamics in online news. Depending on the research perspective, dynamics might be understood as the spreading of a particular data file. Hussain (2012), for example, explores online gatekeeping mechanisms that mediate the distribution of the top 120 viral videos in the 2008 U.S. presidential campaign and finds that relevant actors have expanded well beyond journalists. In a similar take, Trilling et al. (2017) examine factors that contribute to a news article's *shareworthiness* on social media. Here, dynamics refer to the development of an article's success based on its distribution via social media. Based on less clear-cut pieces of information, Leskovec et al. (2009) explore online news dynamics in terms of the handoff of *memes* – distinctive ideas or sentences – through the global news cycle, with results indicating a common pattern of influence from mainstream to social media. In yet another take on online news dynamics, Weber and Monge (2011) trace the information flow by means of hyperlink networks, highlighting the critical role of a few key websites in effectively controlling the overall conversation. Also based on hyperlinks, Messner

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

and Distaso (2008) analyzed the sources cited in weblogs in order to determine who sets their agenda, and the other way around, examined blogs as a reference in mainstream media. The latter two studies consequently define dynamics as an attribute of the online news ecosystem's structure rather than as the flow of particular data or information.

In light of these various approaches, it becomes apparent that analyzing the dynamics in online environments is not an easy undertaking. Journalism researchers have quickly adapted to the new challenges that the digital transformation has afforded. Various perspectives, based on innovative methodological approaches, add to our understanding of digital news as a complex living entity. Each approach has its own advantages. With our case study on online news diffusion, we add a framework that focuses on the speed and shape with which news on real-world events spreads through the online news ecosystem, and, by means of the described methodological walkthrough, hope to provide inspiration for future research.

Further reading

For readers who are interested in research that – in line with our own work (Buhl et al., 2016, 2018) – maps the diffusion of stories in the digital news ecosystem, we recommend Welber's (2016) thesis on *Gatekeeping in the Digital Age*, as well as Harder et al. (2017) "news story" approach to "Intermedia Agenda Setting in the Social Media Age". With regard to data analysis, there are many more ways to explore the content of large text collections. For an overview of various automated content analysis techniques, Günther and Quandt's (2016) "Word Counts and Topic Models" might serve as a good starting point. Alvarez's (2016) collected volume on the *Computational Social Sciences* is recommended for readers with an overall interest in this field, with excellent case studies that might serve as inspiration for future research.

Note

1 Unfortunately, performance optimization was only implemented when the data collection for our case study was already up and running. With the current setup, a much shorter interval is well within reach, e.g., collecting news articles every 15 minutes in future projects.

References

Alvarez, R. M. (ed.). (2016) *Computational Social Science: Discovery and Prediction*. New York, NY: Cambridge University Press.

Boczkowski, P. (2010) *News at Work: Imitation in an Age of Information Abundance*. Chicago, IL: University of Chicago Press.

Buhl, F., Günther, E. and Quandt, T. (2016) "Unambiguous burstiness: Towards explaining the dynamics of digital news flows from opportunity structures, news factors, and topics." Paper presented at the *66th Annual Conference of the International Communication Association*, Fukuoka, 9–13 June.

Buhl, F., Günther, E. and Quandt, T. (2018) "Observing the dynamics of the online news ecosystem: News diffusion processes among German news sites." *Journalism Studies*, 19(1), 79–104.

Donsbach, W. (2004) "Psychology of news decisions: Factors behind journalists' professional behavior." *Journalism*, 5(2), 131–157.

Franklin, B. and Eldridge II, S. (eds.). (2017) *The Routledge Companion to Digital Journalism Studies*. London: Routledge.

Friedl, J. (2006) *Mastering Regular Expressions*. Sebastopol, CA: O'Reilly Media.

Günther, E. and Quandt, T. (2016) "Word Counts and Topic Models: Automated Text Analysis Methods for Digital Journalism Research." *Digital Journalism*, 4(1), 75–88.

Günther, E., Trilling, D. and van de Velde, B. (2018) "But how do we store it? (Big) Data architecture in the social-scientific research process." In C. Stuetzer, M. Welker and M. Egger (eds.), *Computational Social*

Downloaded By: 10.3.98.93 At: 06:51 17 Jan 2019; For: 9781315270449, chapter9, 10.4324/9781315270449-10

Science in the Age of Big Data: Concepts, Methodologies, Tools, and Applications. Cologne: Herbert von Halem Verlag (pp. 161–187).

Harder, R. A., Sevenans, J. and van Aelst, P. (2017) "Intermedia agenda setting in the social media age: How traditional players dominate the news agenda in election times." *The International Journal of Press/Politics*, 22(3), 275–293.

Hussain, M. (2012) "Journalism's digital disconnect: The growth of campaign content and entertainment gatekeepers in viral political information." *Journalism*, 13(8), 1024–1040.

Karlsson, M. and Strömbäck, J. (2010) "Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news." *Journalism Studies*, 11(1), 2–19.

Leskovec, J., Backstrom, L. and Kleinberg, J. (2009) "Meme-tracking and the dynamics of the news cycle." *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 497–506).

Lim, J. (2012) "The mythological status of the immediacy of the most important online news: An analysis of top news flows in diverse online media." *Journalism Studies*, 13(1), 71–89.

Messner, M. and Distaso, M. W. (2008) "The source cycle: How traditional media and weblogs use each other as sources." *Journalism Studies*, 9(3), 447–463.

Risley, F. (2000) "Newspapers and timeliness: The impact of the telegraph and the internet." *American Journalism*, 17(4), 97–103.

Rogers, E. M. (2000) "Reflections on news event diffusion research." *Journalism & Mass Communication Quarterly*, 77(3), 561–576.

Rosenberg, H. and Feldman, C. S. (2008) *No Time to Think: The Menace of Media Speed and the 24-Hour News Cycle*. New York, NY: Continuum International.

Saltzis, K. (2012) "Breaking news online: How news stories are updated and maintained around-the-clock." *Journalism Practice*, 6(5/6), 702–710.

Schütz, W. (2012) "Redaktionelle und verlegerische Struktur der deutschen Tagespresse." *Media Perspektiven*, 11.

Trilling, D. (2017) "Doing computational social science with python: An introduction." *Social Science Research Network*. Retrieved from http://papers.ssrn.com/abstract=2737682

Trilling, D., Tolochko, P. and Burscher, B. (2017) "From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics." *Journalism & Mass Communication Quarterly*, 94(1), 38–60.

Waldherr, A., Maier, D., Miltner, P. and Günther, E. (2017) "Big data, big noise: The challenge of finding issue networks on the web." *Social Science Computer Review*, 35(4), 427–443.

Weber, M. S. and Monge, P. (2011) "The flow of digital news in a network of sources, authorities, and hubs." *Journal of Communication*, 61(6), 1062–1081.

Welbers, K. (2016) *Gatekeeping in the Digital Age*. Doctoral thesis, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. Retrieved from https://research.vu.nl/ws/portalfiles/portal/1587652

Widholm, A. (2016) "Tracing online news in motion: Time and duration in the study of liquid journalism." *Digital Journalism*, 4(1), 24–40.