Toward a better understanding of news user journeys: A markov chain approach

Vermeer, S.; Trilling, D.

DOI
10.1080/1461670X.2020.1722958

Publication date
2020

Document Version
Final published version

Published in
Journalism Studies

License
CC BY

Link to publication

Citation for published version (APA):
Vermeer, S., & Trilling, D. (2020). Toward a better understanding of news user journeys: A markov chain approach. *Journalism Studies*, *21*(7), 879-894.
https://doi.org/10.1080/1461670X.2020.1722958

Routledge
Taylor & Francis Group

Journalism Studies

ISSN: 1461-670X (Print) 1469-9699 (Online) Journal homepage: https://www.tandfonline.com/loi/rjos20

Toward a Better Understanding of News User Journeys: A Markov Chain Approach

Susan Vermeer & Damian Trilling

To cite this article: Susan Vermeer & Damian Trilling (2020) Toward a Better Understanding of News User Journeys: A Markov Chain Approach, Journalism Studies, 21:7, 879-894, DOI: 10.1080/1461670X.2020.1722958

To link to this article: https://doi.org/10.1080/1461670X.2020.1722958

Published online: 13 May 2020.

JOURNALISM STUDIES
2020, VOL. 21, NO. 7, 879–894
https://doi.org/10.1080/1461670X.2020.1722958

OPEN ACCESS Check for updates

Toward a Better Understanding of News User Journeys: A Markov Chain Approach

Susan Vermeer and Damian Trilling

Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT
In recent years, the volume of clickstream and user data collected by news organizations has reached enormous proportions. As a result, news organizations—as well as journalism scholars—face novel methodological challenges to describe and analyze this wealth of information. To move forward, we demonstrate a computational approach to understand the news journeys Web users take to find the news they want to read. We propose the use of Markov chains. These models provide an effective and compact way to discover meaningful patterns in clickstream data. In particular, they capture the sequentiality in news use patterns. We illustrate this approach with an analysis of more than 1 million Web pages, from 175 websites (news websites, search engines, social media), collected over 8 months in 2017/18. The analysis of such data is of high interest to journalism scholars, but can also help news organizations to design sales strategies, provide more personalized content, and find the most effective structure for their website.

KEYWORDS
Online news consumption; Digital journalism; News user journeys; Web usage mining; sequential data; Markov chains

Introduction

In recent years, the very nature of news consumption has changed drastically as people increasingly consume news online (Flaxman, Goel, and Rao 2016). The shift toward online news consumption, however, poses a great challenge to scholars who need to describe and analyze these patterns of news use. For instance, researchers studying how citizens interact with journalistic content more and more find themselves in a position where they need to rely on online tracking data of user behavior (see e.g., Kleppe and Otte 2017; Dvir-Gvirsman, Tsfati, and Menchen-Trevino 2014; Menchen-Trevino and Karr 2012). And, indeed, such data have many advantages. They tackle the problem of overreporting in surveys on news use (see Prior 2009), and allow for a much more fine-grained level of analysis. For instance, it is possible to go beyond analyzing which outlet has been used, but to examine which articles exactly have been used. This is especially important when researchers want to link audience data to content data in order to study media effects (Scharkow and Bachl 2017; De Vreese et al. 2017).

While such data are of core interest for scholars studying online journalism, it is still difficult to make sense of the wealth of such data. One could, for instance, aggregate all

CONTACT Susan Vermeer S.A.M.Vermeer@uva.nl

content per user to compare the different media diets that users have; or one could study the overlap of sources used (e.g., Mukerjee, Majo-Vazquez, and Gonzalez-Bailon 2018), or one could even construct and analyze a network of users and specific articles (Trilling 2019).

In this article, we want to focus on another, frequently neglected, dimension, namely the *sequence* of news-related pages and sites a user is visiting. We know from peoplemeter data that television viewership is at least partly shaped by lead-in and lead-out effects: watching a given program can make it more likely to watch the program that is aired next or before (Wonneberger, Schoenbach, and van Meurs 2012). Even though the Web user has much more freedom to choose whatever they please than the television viewer has, the hyperlinked nature of the Web suggests that also here, some "default" paths in a users' online news journey emerge. And indeed, Möller et al. (2019) illustrate that Web users can be distinguished based on patterns and modes of news use online (e.g., using search engines to access news online).

But while it is relatively straightforward to sketch such a journey for one given user, it is more difficult to abstract from here and generalize towards typical journeys. To fill this gap, we aim to show what journalism scholars can learn from adapting methods currently in use in the scholarly field of data mining. More specifically, we suggest a technique to do so, which is based on Markov chains, a method (often used in Web usage mining) to conceptualize the probability to move—in our case—from one type of news topic to another one (e.g., politics, entertainment, business). We will focus on the following question: "*How can Markov chains be applied to the field of journalism studies as a way to understand the sequentiality of news use (i.e., news journeys Web users take to find the news they want to read)*?" Our objective is to familiarize journalism scholars with the basic principles of Markov chains and show how these principles can be implemented. We therefore show the application to an exemplary use case.

By means of this approach, we hope to offer substantial added value to journalism research. Our proposed approach aims to help scholars analyzing and understanding (1) relationships between content types, and (2) user behavior. Besides, analyzing how Web users navigate today's media landscape is especially interesting for the journalistic field. News organizations aim to understand the journeys Web users take to find the news they want to read, and how they navigate the complex wealth of online information. By utilizing Markov chains, news organizations (including journalists and Web developers) can better understand news users journeys; and, eventually, can help shape the paths Web users take. We conclude with suggestions for further extending and improving such an approach to overcome some of its limitations.

The Sequentiality of News Use

News Sequences in Journalism Studies

Traditionally, journalism studies and audience research did not pay much attention to the sequence in which news items are consumed. With the exception of analyses of television meter data, which have shown when watching a specific program can increase the chance of watching the preceding or following program, leading to lead-in and lead-out effects (Wonneberger, Schoenbach, and van Meurs 2012), the order in which news items are consumed has not been of central interest. This changed with a development that has been referred to as "unbundling" (Hermida et al. 2012; Trilling 2019), "debundling" (Schmidt et al. 2019) or "atomization" (Bruns 2018) of news: A specific news item is no longer (only) part of a specific daily edition of a newspaper, or a specific broadcast of a specific television news program, but can be encountered in various places through various routes. Early theories of online journalism emphasized its non-linearity and pointed to the role of hyperlinks as a means for users to interactively determine their own path through, the possible news content (e.g., Deuze 2003). Yet, for a long time, the actual use of hyperlinks on news sites remained much more limited than early theories expected (see, e.g., Barnhurst 2010).

More recently, the focus shifted more towards the role of intermediaries, such as social networking sites. For instance, Thorson and Wells (2016) have introduced a concept of "curated flows", in which they argue that the path a user takes to arrive at a given news item is shaped by the gatekeeping logic (curation logic) of journalists, algorithms, and peers. Yet, also studies following such approaches usually measure outcomes such as the number of articles exposed to rather than the *sequence* of articles visited (e.g., Wells and Thorson 2017; Fletcher and Nielsen 2018). In fact, in neighboring fields, such as the study of (news) recommender systems, sequences play an important role: For instance, some studies investigate whether or not recommender systems draw users into a rabbit hole of increasingly extreme content, such as O'Callaghan et al. (2015) who argue that YouTube's suggestions what to watch next tend to be more extreme than what has been watched before.

We argue that to fully utilize the theoretical potential of approaches that conceptualize news usage as a dynamic process, a flow, it is necessary to model exposure to individual news items as dependent on what has been watched or read before. We thus need to model users' journey through a set of potential news items.

While some studies have used network analysis to understand clickstreams (Taneja, Wu, and Edgerly 2018), we offer an alternative approach in which we are able to detect trends or movements, which cannot be found from a single news article but from a sequence of articles. In particular, our approach offers a very easy to understand interpretation, as each relation between two states is simply the probability of transitioning from one state to the other.

Although the information on news websites can be understood as a sequence of news articles sets, it can be examined from various perspectives. In the example we provide in this article, we focus on the probability of users changing from one news topic to another news topic. By doing so, we argue that the news topic of a current news article is expected to be able to grasp the news topic of the next news article. In a high-choice media environment, political topics constantly compete with entertaining topics (Prior 2005); as many people generally prefer the latter (Prior 2009). Since both entertainment and political news are available on numerous websites, people's content preferences become key to understanding online news consumption (Prior 2005).

Besides examining the probability of users changing from one news topic to another news topic, it is also possible to detect a sequence in clickstream data by examining the probability of users changing from one website to another website (e.g., *social media → tabloid → tabloid → broadsheet)*, which could help to understand the differences in news consumption patterns contributed by each of the different (news) websites

(Kleppe and Otte 2017). Another possibility would be to examine the probability of users changing from one Web page to another Web page within the same website (e.g., *homepage → section page → news article → news article*). This type of analysis could provide relevant insights for news organizations seeking to guide users to relevant content on their website (e.g., enabling personalization). As journalists, editors and Web developers are the architects of news pathways (Pearson and Kosicki 2017), such findings could help them shape the paths users take. And, as we have seen, especially from a curated-flows perspective Thorson and Wells (2016), understanding such paths is of crucial importance for contemporary journalism scholarship.

In any event, it is safe to say that, even though journalism studies acknowledges the importance of sequential news journeys in today's media environment, these theoretical considerations have rarely been subjected to empirical analyses.

News Sequences in Journalism Practice

Not only in journalism studies, but also in journalism practice, the sequentiality of news gains importance.

First, journalistic news production is more and more metrics-driven, and paying attention to audience metrics and the question which stories are clicked on most have become ingrained into newsrooms routines (Tandoc 2014; Giomelakis et al. 2019). A/B testing (or split testing), in which different versions of headlines are used for the same article in order to determine which one generates the most clicks, is commonplace in contemporary journalism as well (Hagar and Diakopoulos 2019). But arguably, the reason to read an article does not only depend on the features of that article, but also on where the user comes from—in other words, sequence matters.

Second, newsrooms started experimenting with the use of recommender systems to give personalized recommendations for their readers (Sappelli et al. 2018; Möller et al. 2018). While the algorithms that are implemented to this end differ a lot in their complexity, one could argue that at least conceptually, it makes sense to distinguish between an article that the reader has read immediately before the current article and an article that the reader has read at one arbitrary point in time.

In order to get a better idea of the state of the art of the use of such approaches, and to be able to better put into perspective our proposed approach, we aimed to obtain insights about the role of data for national, regional and local news brands. To get such background information, we interviewed a senior data analytics manager, responsible for generating insights for several Dutch and Flemish news brands. He highlighted that the volume of clickstream and user data has reached enormous proportions:

> Colleagues using the data (e.g., editorial groups, product owners) absolutely drown in the amount of available data. There are many possibilities to analyze the data, making it fun and risky at the same time.

He also pointed out that access to a wide variety of data sources does not necessarily indicate that an organization is considered to be at a "mature" stage of data analysis. Knowing how many daily visitors the website has and how many visitors subscribe to the news brand is simply not enough to know how to improve digital products. This requires different data:

> 'What type of stories do visitors read?', 'What type of stories make it that visitors opt-in or opt-out?'. Such insights provide us with better predictions whether visitors return to our website in the future. Frequently returning to our website increases levels of loyalty, and, subsequently, the chance they will convert. (i.e., website visitors turning into paying customers)

To reach such goals, news brands focus on two different concepts, namely:

(1) format (e.g., by Web developers and product owners), to improve the website and reach a high "read-depth" (i.e., visitors reading an entire article) and "page-depth" (i.e., visitors click-through another news article). And, (2) content (e.g., by editorial groups), to predict whether visitors opt-out on certain news articles or topics, or decide to subscribe to the news brand. During this process, the sequentiality of news is important:

> The sequence is definitely of importance, it is part of the customer journey: 'What are visitors doing during a Web session?', 'Are visitors interested in a wide variety of topics? Or are they reading a specific set of articles?'. All in all: 'Are visitors using the entire spectrum that we offer?' (…) Particularly, if visitors are exposed to elements both of the news section and the service section, there is an increased chance that they will convert. By trying to understand user behavior—not by asking, but by examining—we are able to align our digital products to users' needs, both in terms of format and content.

Discovery and Analysis of Web Usage Patterns

News consumers, thus, provide journalists insightful information by merely clicking on a news item. These records, which are logged in log files, can serve as a source of information on *which* news items the audience is interested in. Since the volume of clickstream and user data collected by news organizations has reached enormous proportions, news organizations—as well as scholars—face new challenges to describe and analyze this wealth of information. According to Zamith and Lewis (2015) this points to a growing recognition of the benefits of using computational and algorithmic solutions to drive analyses.

Analyzing such data would help news organizations, among other things, to design sales strategies, provide more personalized content, and find the most effective structure for their website. This type of analysis is often referred to as *Web usage mining*: The automatic discovery and analysis of meaningful patterns and relationships from a large collection of semi-structured data, such as clickstream data (e.g., log files; Mobasher 2007).

The Web usage mining process involves three different stages: (1) data collection and pre-processing (e.g., processing clickstream data into a sets of user activities), (2) pattern discovery (e.g., obtaining hidden patterns reflecting patterns of Web usage), and (3) pattern analysis (e.g., building user models). In this paper, we particularly focus on the third stage, namely the analysis of Web usage patterns.

Various types of analysis can be used to analyze sequential Web usage data (also depending on the desired outcomes; for an overview see Mobasher 2007). Session and user analysis have often been used to analyze pre-processed clickstreams. News managers consult this type of data to see the number of unique users at their website, the most frequent accessed news items, the average view time of a news item, and the most common entry- and exit-points (MacGregor 2007). In a more detailed way, journalists can use this type of analysis to discover how a news item—be it a single news item, a Web page, or an entire news section—has performed or is performing with the audience. For example, Boczkowski, Mitchelstein, and Walter (2011) compared online news choices of journalists (the most prominently shown items on news websites' homepages) and news consumers (the most frequently clicked news items). The results indicated a gap in the online choices of journalists and news consumers: journalists were more likely to select hard news topics (e.g., international affairs, politics, economics) as the most newsworthy stories, whereas consumers' choices were dominated by soft news topics (e.g., sports, entertainment, crime). Similarly, Tewksbury (2003) found that online news consumers do not select hard news content as often as they select other news content. Despite a lack of depth, this type of analysis can be potentially useful for, for example, understanding users' interests.

Additionally, cluster analysis and user segmentation can be used to analyze Web usage data. Clustering is a data mining technique that groups together a set of items (or individuals) having similar characteristics. Clustering of users tends to establish groups of users showing similar patterns of Web usage. For example, in a recent study Möller et al. (2019) used *k*-means clustering to analyze online tracking data. The results illustrate four relatively distinct clusters of news users, for example, distinguishing users who access news items through search engines and users who access news items (more incidentally) through social media. The results of this type of analysis can be potentially useful for providing personalized Web content to users with similar interests.

Markov Chains

A comparatively easy and straightforward way to model news user journeys are Markov chains. Markov chains have already been successfully applied in a wide range of domains including economics and finance (e.g., predicting asset prices; Tauchen 1986), sports (e.g., baseball analysis; Bukiet, Harold, and Palacios 1997), games (e.g., Snakes and Ladders; Althoen, King, and Schilling 1993), mathematical biology (e.g., simulations of brain function; George and Hawkins 2009), search-engine algorithms (e.g., Google PageRank algorithm; Rai and Lal 2016), and speech recognition (Juang and Rabiner 1991). Besides, smartphone's predictive text feature relies on the idea behind Markov chains: predicting what a user is going to type next, based on the last thing they typed (i.e., the sequencing of neighboring words). Essentially, the task of predicting a word a user might want to type given the last word they typed does not differ much from the task to predict which type of article a user might want to read given the article read last.

Despite their mathematical simplicity, Markov chains have not yet been widely implemented in communication research. Only recently, Hopp, Fisher, and Weber (2019) used hidden Markov models to predict news frames and real-world events sequences in the United States; and, eventually, forecasting trends in news frames and event densities (e.g., political and civil conflicts surrounding gun violence, the Trump administration family separation policy). Distinct event occurrences might be followed by certain news frames which in turn contribute to the onset of future events. They not only highlight the feasibility of utilizing Markov models, but also its future potential for advancing communication research. Besides for journalism scholars, Markov chains could also be highly relevant for sequential data of mobile communication scholars (e.g., analyzing transitions between mobile application usages), marketing communication scholars (e.g., analyzing consumer

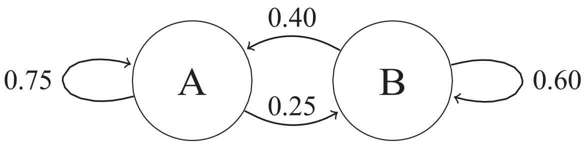

Figure 1. Example of a two-state Markov chain.

behavior), and health communication scholars (e.g., examining usage of healthcare information technologies).

Indeed, as discussed before, also in the field of Web usage mining, several useful and well-studies models have been proposed to discover and analyze patterns in sequential data (see Mobasher 2007). Among the proposed methods we advocate the use of Markov chains. Markov chains provide an effective and compact way to represent Web usage data, and are especially suited to detect Web pages that are often viewed in a sequence (i.e., sequential pattern mining). Markov chains are mathematical systems that change from one state (i.e., a situation) to another state (Fygenson 1989). All possible states are listed in a *state space* $\{s_1, s_2, \dots, s_n\}$. A Markov chain provides the likelihood of transitioning from state s_i to state s_j. The probabilities for all transitions are presented in a transition probability matrix: $[\mathrm{Pr}_{i,j}]_{n\times n}$, where $L_{j=1}^{n}p_{i,j} = 1$ for every i.

Figure 1 illustrates an exemplary two-state Markov process with two states (A and B). In total, there are four possible transitions, as a state can transition back into itself. Each number represents the probability of the Markov process changing from one state to another state, with the direction indicated by the arrow (Gilks, Richardson, and Spiegelhalter 1996). For example, if the Markov process is in state A, then the probability it changes to state B is 0.25, whereas the probability it remains in state A is 0.75.

Markov chains offer a high flexibility regarding the level of analysis. We can construct a Markov chain for an individual user to model their individual probability for transitioning from one state to another. One could then either analyze this individual's behavior in-depth, or compare the distribution of transition probabilities across users. But we can also simply construct a Markov chain on an aggregated level, neglecting the individual user. By calculating transition probabilities across an entire collection of Web user sessions at an aggregate level. In this way, Markov chains are not only able to predict the *next* user action based on a users' previous surfing behavior, but also discover high probability sequences (e.g., important relationships among news items), and perform exploratory analysis of overall user behavior within or across different websites.

Despite their mathematical simplicity, Markov chains have not yet been widely implemented by journalism scholars. In the next section, we therefore demonstrate how patterns of Web usage can be modeled as a Markov chain.

Application to an Exemplary Use Case

In the following section, we will show how this method can be put into practice. To do so, we will walk the reader through all steps that need to be done in order to analyze a journey. This does not require any specific software. In essence, it only requires counting transitions and calculating percentages. To encourage other scholars to use Markov chains, we have developed a Python module, *df2markov*, that automatically performs all necessary calculations and can be used by any scholar for their own data sets.[1]

Working Example

In this paper, we demonstrate how Web usage mining, in particular Markov chains, can be applied to analyze online tracking data, as tracking online Web users offers a new avenue for journalism scholars. To do so, we used online tracking data collected by a joint Communication and Information Law Initiative of the University of Amsterdam.

In this project, 383 participants[2] installed a browser plug-in *Robin*—a custom-built system that registers participants' Web behavior—and had their browsing behavior being monitored in the period between 1 July 2017 and 15 March 2018. Participants were recruited via the LISS (Longitudinal Internet Studies for the Social Sciences) panel of CentERdata; a true probability sample of the Dutch population. Whenever participants accessed one of the white-listed websites (which includes an exhaustive list of various websites: news websites as well as social media and search engines), the plug-in transmitted all Web traffic (HTTP and HTTPS) to the servers.[3] In this way, the system collected the URL request (Web page address), the date, time, session ID and referrer URL (i.e., the address of the Web page that linked to the resource being requested) for each recorded visit (see Table 1). In total, we collected 2,336,164 URLs.

Analysis

To be able to analyze participants' online activities, the raw data needs to be captured and transformed into relevant and meaningful information:

(1) Online tracking data: During the first stage of the pre-processing phase, it is important to decide the scope of the study. In this working example, Web behavior is extracted for 175 websites. In our article, a news website refers to a Web page whose primary offering is news content, varying from tabloids (e.g., www.telegraaf.nl, www.ad.nl), broadsheets (e.g., www.volkskrant.nl, www.nrc.nl), online-only outlets (e.g., www.nu.nl), international outlets (e.g., www.dailymail.co.uk, www.lemonde.fr), broadcasters (e.g., www.nos.nl, www.npo3.nl), regional outlets (e.g., www.parool.nl, www.at5.nl), and other news websites (e.g., www.metronieuws.nl). Besides, we included search engines (e.g., www.google.nl, www.yahoo.com) and social media (e.g., Facebook, Twitter) to get a more complete picture of online news patterns. During the first

Table 1. An exemplary selection of the data from user Z.

Date	Time	User	URL
08/09/18	11:44:04	Z	https://www.google.nl
08/09/18	11:45:49	Z	https://www.ad.nl/binnenland/nederlanders-bewegen-meer-dan-wie-ook-in-de-eu/
08/09/18	11:49:09	Z	https://www.ad.nl/buitenland/
08/09/18	11:49:51	Z	https://www.volkskrant.nl/nieuws-achtergrond/eu-voorzitter-verwacht-snel-akkoord-digitaks/
08/09/18	11:52:37	Z	https://www.volkskrant.nl/nieuws-achtergrond/het-binnenhof-legt-de-mantel-der-liefde-af/
08/09/18	11:53:29	Z	https://www.volkskrant.nl/nieuws-achtergrond/weerbericht-veelal-droog-weekeinde/
08/09/18	12:49:16	Z	https://www.weeronline.nl

pre-processing step, we removed irrelevant URLs (e.g., marketing Web pages of news websites; www.abonnement.parool.nl), which resulted in a sample of 1,175,022 URLs by 354 unique users.

(2) Web session: Next, we determine what we consider a Web session: for example, a user can click on a news item, and after having read the article, proceed to click links to other articles published on the news website. To define new sessions, we implement the following algorithm: (a) group the tracking data by user ID and time, (b) create a new session if there is a 30-minute gap between records in the data (using the same methodology as Athey and Mobius 2012), and (c) assign the same session ID to records that are connected. For example, in Table 1, we would assign a session ID to the first six records, and start a new session for the final record in the list.

(3) State space: Next, we define a *state space*, which lists all possible states. In this example, we will focus on the probability of users changing from one news topic to another news topic. Though, as we mentioned before, there are numerous other possibilities that can be explored using Markov chains.[4] In this working example, a supervised machine learning approach (Vermeer 2018) was used to classify news items. The coding scheme developed by Shoemaker and Cohen (2005) was used to guide the coding process. Four news categories were distinguished: (1) Politics (e.g., *internal politics, international politics, and military and defense*); (2) Business (e.g., *economy, and education*); (3) Entertainment (e.g., *sports, culture, and human interest)*; and (4) Other (e.g., *science and technology, environment, and religion and beliefs)*. To classify the topic of news items,[5] we assign one of the following *states* to every record in our data: (1) Homepage or section page, (2) Politics, (3) Business, (4) Entertainment, (5) Other, and (6) End of Web session. The final "absorbing" state (i.e., End of Web session) is added to every session representing the exit point (i.e., a state once entered, cannot be left). In total, we distinguish six states (Vermeer 2018).

(4) Sequential patterns: Subsequently, we have to organize the data into meaningful sequential patterns. By grouping the data by user ID, time and session ID, a session can be defined as a sequence of states. We are now able to create a transition probability matrix: the likelihood of transitioning between any two states. In total, there are six states, which results in a transition matrix of 6×6. In a probability transition matrix, row *i* contains the transition probability from s_i to any other state s_n, and always needs to sum up to 1 (Mobasher 2007).

(5) Evaluation: Before making predictions using the Markov models, we must evaluate the reliability of the estimated transition matrices.[6] This is less straightforward than it seems. It lies at the very nature of our Markov model that for any given user at any given stage, we cannot say what they will do next *in this instance*. We can only say what are—in the long run—the proportions of each possible state the user will turn to. Imagine a hypothetical two-state Markov model that perfectly fits reality and that has the transition probabilities depicted in Figure 1. If we would use this model to predict transitions for new users later on, and would for every user who is in state A predict that they would stay in A (because $0.75 > 0.25$), then even though this would be the most reasonable prediction, we would expect to be wrong in 25% of the cases. Using classical machine learning evaluation metrics like accuracy, precision, and recall therefore are not helpful in our case. Instead, the predictive accuracy of our model, in this example, would be perfect if also in a new data set, the

probability $A \rightarrow A$ turns out to be .75. In this paper, we stick to the simple case of Markov chains without "memory", i.e., for any prediction, we only take into account the current state (and not previous states); if, in contrast, we would use higher-order Markov chains that do that, we would have several similar metrics at our disposal that could quantify how well we model complete trajectories of users (see, e.g., Eirinaki, Vazirgiannis, and Kapogiannis 2005). For the sake of this paper, however, we use a simpler approach. We applied a random sampling procedure to split the data set into a training set of Web sessions (80%), on which we trained the Markov model, and a test set (20%), on which we evaluated the model. Various statistical tests are available that we could use to evaluate the model, such as a likelihood ratio statistic and the Pearson chi-square statistic (see e.g., Billingsley 1961; Hiscott 1981; Anderson and Goodman 1957). In our working example, we employ a Kolmogorov-Smirnov goodness-of-fit test (like Lazariv and Lehmann 2018), as we are dealing with a large sample size. As Table 3 shows, the probabilities observed in the test data set are almost identical to those that we would have predicted based on our model depicted in Table 2, which is confirmed by the K-S goodness-of-fit test (Tables 2 and 3).

(6) Markov chain: Finally, we capture sequential dependence by modeling users' navigational behaviors. Based on the transition probability matrix, we define the transition paths (as presented in Figure 1) at an individual-level as well as at an aggregate-level.

Figure 2 represents any possible transition in terms of news items' content. In this way, we argue that the topic of the current news article is expected to be a good clue to grasp the topic of the next news article. We assigned weights to the edges to represent the probability of users changing from one news topic to another news topic. In other words, the thicker the line, the higher the transition probability. In this graph, we can discover two clear sequential patterns. First, after reading a news item (about any topic) users frequently return to a homepage or section page (H) and continue browsing from there. Second, there is a relative preference for entertainment news. Online news consumers are very likely to transfer to entertainment news during a Web session, and also continue reading entertainment news.

Using the transition paths, we are now able to implement the Markov model to predict the next click, or the state after n clicks. For example, in order to provide recommendation (either personal or not), it is possible to choose the path with higher probability among all existing paths (see Figure 2).

Table 2. Transition probabilities training set (80%)—six states.

	Homepage	Politics	Business	Entertainment	Other	End of Web session
Homepage	15.99	4.27	26.16	49.92	.79	2.88
Politics	14.09	5.69	27.91	47.65	.81	2.33
Business	12.60	4.01	37.16	41.71	1.06	3.45
Entertainment	14.16	4.13	23.71	54.08	.	.
Other	11.70	6.46	23.84	49.68	2.61	2.70
End of Web session	.	.	.	.	.	100

Notes: K-S goodness-of-fit test: $D(N_{sessions} = 348{,}312) = 0.07$, $p = .99$.

Table 3. Transition probabilities test set (20%)—six states.

	Homepage	Politics	Business	Entertainment	Other	End of Web session
Homepage	15.23	3.50	23.86	50.45	.70	4.75
Politics	14.75	4.50	25.47	45.95	1.14	2.13
Business	13.40	4.35	35.38	38.66	.94	2.71
Entertainment	13.65	4.19	21.41	54.50	.	.
Other	10.41	3.86	21.39	42.33	4.08	1.26
End of Web session	.	.	.	.	.	100

Notes: K-S goodness-of-fit test: $D(N_{sessions} = 348{,}312) = 0.07$, $p = .99$.

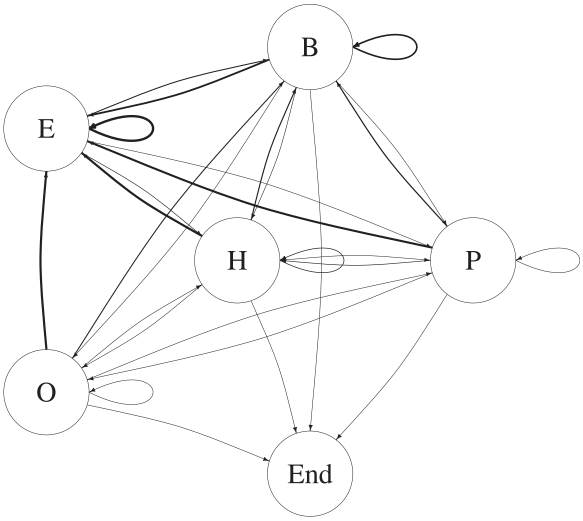

Figure 2. The probability of users changing from one news topic to another news topic: Homepage or section page, Politics, Business, Entertainment, Other, and **End** of Web session.

Additionally, it is important to take into account how the state space— defined at step three—affects the predictive capabilities of a Markov chain. To demonstrate how this can affect the interpretation, we merged “Politics” and “Business” news topics to a *hard* news category, and “Entertainment” and “Other” news topics to a *soft* news category. Comparing 3 with 2, we see that this new model still leads to similar conclusions regarding the dominance of soft news, but the really low probabilities towards politics are not captured and overshadowed by the comparatively well-doing business news. As this example illustrates, using too broad categories to define the state space can lead to less informative models, even if they fit the data well (Figure 3; Tables 4 and 5).

Conclusion and Discussion

In this paper, we have argued that to understand online news use, we must take into account the sequentiality of users’ journey. We need to move beyond analyzing why a user clicked on a given news item, and ask how the user actually got there. In particular, we suggested a simple to use yet informative model: Markov chains, which describe the probability of transitioning from one *state* (such as “reading an entertainment article”,

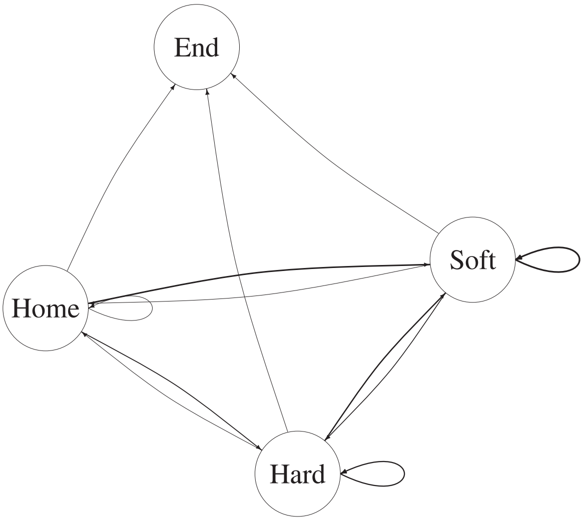

Figure 3. The probability of users changing between hard and soft news: **Home**page or section page, **Hard** news, **Soft** news, and **End** of Web session.

Table 4. Transition probabilities training set (80%)—four states

	Homepage	Hard news	Soft news	End of Web session
Homepage	15.60	30.56	50.40	3.44
Hard news	13.01	39.67	44.07	3.24
Soft news	14.62	27.70	54.92	2.76
End of Web session	.	.	.	100

Notes: K-S goodness-of-fit test: $D(N_{sessions} = 348{,}312) = 0.17$, $p = .99$.

Table 5. Transition probabilities test set (20%)—four states.

	Homepage	Hard news	Soft news	End of Web session
Homepage	15.43	30.72	50.09	3.76
Hard news	13.80	38.83	43.95	3.41
Soft news	13.86	26.43	55.90	3.82
End of Web session	.	.	.	100

Notes: K-S goodness-of-fit test: $D(N_{sessions} = 348{,}312) = 0.17$, $p = .99$.

or “being on Facebook”) to another *state* (such as “reading a politics article”, or “being on the publisher’s website”). The states of interest can be anything, and have to be defined in advance, forming the *state space*.

This is interesting both for journalism studies and journalism practice. Although modeling Markov chains is increasingly popular in Web usage mining, there are only a limited number of studies using Markov chains in communication research (see e.g., Hopp, Fisher, and Weber 2019). To fill this gap, the present article explicates the usage of Markov chains in the communication field. We aim to make a methodological contribution to journalism research, addressing earlier calls to adopt computational and algorithmic solutions in the field (Zamith and Lewis 2015). Since the volume of clickstream and user data has reached enormous proportions, scholars face new challenges to describe and analyze this wealth of information. By using Markov chains, we demonstrate an effective approach to discover meaningful patterns from clickstream data. In this way, we are able to obtain a further understanding of the news journeys Web users take to find the news they want to read.

Describing and analyzing patterns of news use is also essential for journalism practice. In particular, the sequentiality of news has gained importance. News user journeys can be an important source of support and inspiration for editorial groups. Moreover, such insights help to determine whether a certain news article should be offered merely to premium members or not. And, are helpful to develop and improve personalization and content recommendation. To quote the data analyst from the publishing house again:

> This next step is still rather difficult: 'What do we offer visitors after reading a news article?' If they read a political news article, we often present them other political news articles to increase recirculation. Though, you have to offer different interesting news articles that people do not expect, to prevent people from entering a certain funnel.

The shift toward online news consumption has improved the ability to describe and analyze patterns of news use. Although novel methods for the study of online news consumption have opened up, it remains challenging:

> This is a search (…). Last week, when I was talking to the editor- in-chief, I compared it to peeling off an onion. In the past we did not go beyond the first two layers, and now we are getting closer and closer to the core, but we are definitely not there yet …

In this paper, we have described a very basic Markov chain model. We deliberately chose to do so, because we think that its parsimonity lowers the threshold for applying it substantially. In particular, it has the great advantage of being very straightforward to interpret. Even people without an in-depth understanding of the methodology can readily interpret the meaning of the probabilities in Figure 1 or Figure 2. This makes our method also suitable for the communication of applied research.

This simplicity, however, comes at a cost. Most notably, the model we presented is "memory-less": we only consider the last state in which a user is to predict the next state. We do not consider, for instance, whether they already been at the next state before; or what they did two steps ago. Utilizing "memory-less" Markov chains to explore sequential news use, scholars should explore how often a meaningful journey (Topic → Topic → Topic) is interrupted by visiting the homepage (Topic → Homepage → Topic). In our data set, the latter occurred for approximately 10% of the transitions. Future work should explore the application of Markov chains that do have such a memory. While full Markov chains that take the whole history of a process into account may be not only inefficient but also unnecessary, the use of variable length Markov chains (VLMC) might be worth exploring (see, for instance, Ferrari and Wyner 2003; Machler and Buhlmann 2004).

In conclusion, Markov chains can be effectively used to (1) analyze individual user behavior in-depth, and compare, for example, the distribution of transition probabilities across users, and (2) analyze user behavior at an aggregate level to discover high probability sequences (e.g., important relationships among news items). This, in turn, enables news organizations as well as research to obtain a better understanding of news user journeys within or across different websites.

Notes

1. More information on the Python module see https://github.com/uvacw/df2markov.
2. Besides tracking their online media use, respondents also filled out an online survey: 48.5% were male, mean age was 47.2 (SD = 19.2), and 15.7% had a low level of education (e.g., primary school), 38.3% had a medium level of education (e.g., college), and 44.6% had a high level of education (e.g., university).
3. To guarantee respondents' privacy as much as possible, we filtered the raw data to exclude sensitive information. We stored the data in an Elasticsearch database on a server that is not directly available for the researchers. Instead, *Robout*, a Python library is made available on another secured server to complement *Robin*. We conducted the analyses using *Robout* and a Elasticsearch database on the second server so no sensitive data would leave the environment.
4. Examining the probability of users changing from one website to another website (e.g., *social media* → *tabloid* → *tabloid* → *broadsheet*) or the probability of users changing from one Web page to another Web page within the same website (e.g., *homepage* → *section page* → *news article* → *news article*).
5. More information see https://doi.org/10.6084/m9.figshare.7314896.v1.
6. We are grateful to an anonymous reviewer for their suggestion.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The research was supported by the Research Priority Area "Personalised Communication" of the University of Amsterdam. The work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

ORCID

Susan Vermeer http://orcid.org/0000-0002-9829-8057
Damian Trilling http://orcid.org/0000-0002-2586-0352

References

Althoen, S. C., L. King, and K. Schilling. 1993. "How Long Is a Game of Snakes and Ladders?" *The Mathematical Gazette* 77 (478): 71–76.

Anderson, T. W., and Leo A. Goodman. 1957. "Statistical Inference About Markov Chains." *The Annals of Mathematical Statistics* 28 (1): 89–110.

Athey, Susan, and Markus Mobius. 2012. "The Impact of News Aggregators on Internet News Consumption: The Case of Localization."

Barnhurst, Kevin G. 2010. "The Form of Reports on Us Newspaper Internet Sites, an Update." *Journalism Studies* 11 (4): 555–566.

Billingsley, Patrick. 1961. "Statistical Methods in Markov Chains." *The Annals of Mathematical Statistics* 32 (1): 12–40.

Boczkowski, Pablo J., Eugenia Mitchelstein, and Martin Walter. 2011. "Convergence across Divergence: Understanding the Gap in the Online News Choices of Journalists and Consumers in Western Europe and Latin America." *Communication Research* 38 (3): 376–396.

Bruns, Axel. 2018. *Gatewatching and News Curation: Journalism, Social Media, and the Public Sphere*. New York, NY: Lang.

Bukiet, Bruce, Elliotte Rusty Harold, and Jose Luis Palacios. 1997. “A Markov Chain Approach to Baseball.” *Operations Research* 45 (1): 14–23.

De Vreese, Claes H., Mark Boukes, Andreas Schuck, Rens Vliegenthart, Linda Bos, and Yph Lelkes. 2017. “Linking Survey and Media Content Data: Opportunities, Considerations, and Pitfalls.” *Communication Methods and Measures* 11 (4): 221–244.

Deuze, Mark. 2003. “The Web and Its Journalisms: Considering the Consequences of Different Types of Newsmedia Online.” *New Media & Society* 5 (2): 203–230.

Dvir-Gvirsman, S., Y. Tsfati, and E. Menchen-Trevino. 2014. “The Extent and Nature of Ideological Selective Exposure Online: Combining Survey Responses with Actual Web Log Data from the 2013 Israeli Elections.” *New Media & Society* 18 (5): 857–877.

Eirinaki, Magdalini, Michalis Vazirgiannis, and Dimitris Kapogiannis. 2005. “Web Path Recommendations Based on Page Ranking and Markov Models.” In Proceedings of the seventh ACM international workshop on Web information and data management – WIDM ‘05, New York, NY, USA, pp. 2. ACM Press.

Ferrari, Fiorenzo, and Abraham Wyner. 2003. “Estimation of General Stationary Processes by Variable Length Markov Chains.” *Scandinavian Journal of Statistics* 30 (3): 459–480.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016. “Filter Bubbles, Echo Chambers, and Online News Consumption.” *Public Opinion Quarterly* 80: 298–320.

Fletcher, Richard, and Rasmus Kleis Nielsen. 2018. “Are People Incidentally Exposed to News on Social Media? A Comparative Analysis.” *New Media & Society* 20 (7): 2450–2468.

Fygenson, Mendel. 1989. “A Fundamental Matrix for Regular Semi-Markov Processes.” *Stochastic Processes and Their Applications* 32 (1): 151–160.

George, Dileep, and Jeff Hawkins. 2009. “Towards a Mathematical Theory of Cortical Micro-Circuits.” *PLoS Computational Biology* 5 (10): 1–26.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. New York, NY: Chapman and Hall.

Giomelakis, Dimitrios, Efstathios Sidiropoulos, Sotiria Gilou, and Andreas Veglis. 2019. “The Utilization of Web Analytics in Online Greek Journalism.” *Journalism Studies* 20 (5): 609–630.

Hagar, Nick, and Nicholas Diakopoulos. 2019. “Optimizing Content with A/B Headline Testing: Changing Newsroom Practices.” *Media and Communication* 7 (1): 117.

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

Hiscott, Richard N. 1981. “Chi-Square Tests for Markov Chain Analysis.” *Journal of the International Association for Mathematical Geology* 13 (1): 69–80.

Hopp, Frederic, Jacob Fisher, and Rene Weber. 2019. “The Dynamic Relationship between News Frames and Real-World Events: A Hidden Markov Model Approach.” Paper presented at the annual meeting of the International Communication Association (ICA), Washington, DC, USA.

Juang, Biing Hwang, and Laurence R Rabiner. 1991. “Hidden Markov Models for Speech Recognition.” *Technometrics* 33 (3): 251–272.

Kleppe, Martijn, and Marco Otte. 2017. “Analysing and Understanding News Consumption Patterns by Tracking Online User Behaviour with a Multimodal Research Design.” *Digital Scholarship in the Humanities* 32 (2): 158–170.

Lazariv, Taras, and Christoph Lehmann. 2018. “Goodness-of-Fit Tests for Large Datasets.”

MacGregor, Phil. 2007. “Tracking the Online Audience.” *Journalism Studies* 8 (2): 280–298.

Machler, Martin, and Peter Buhlmann. 2004. “Variable Length Markov Chains: Methodology, Computing, and Software.” *Journal of Computational and Graphical Statistics* 13 (2): 435–455.

Menchen-Trevino, Ericka, and Chris Karr. 2012. “Researching Real-World Web Use with Roxy: Collecting Observational Web Data with Informed Consent.” *Journal of Information Technology & Politics* 9 (3): 254–268.

Mobasher, B. 2007. “Web Usage Mining.” In *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, edited by B. Liu, 450–483. New York: Springer.

Möller, Judith, Damian Trilling, Natali Helberger, and Bram van Es. 2018. “Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and Their Impact on Content Diversity.” *Information, Communication & Society* 7: 1–19.

Möller, Judith, Robbert Nicolai van de Velde, Lisa Merten, and Cornelius Puschmann. 2019. “Explaining Online News Engagement Based on Browsing Behavior: Creatures of Habit?” *Social Science Computer Review*, Advance online publication.

Mukerjee, Subhayan, Slvia Majo-Vazquez, and Sandra Gonzalez-Bailon. 2018. “Networks of Audience Overlap in the Consumption of Digital News.” *Journal of Communication* 68 (1): 26–50.

O’Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Padraig Cunningham. 2015. “Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems.” *Social Science Computer Review* 33 (4): 459–478.

Pearson, George D. H., and Gerald M. Kosicki. 2017. “How Way-Finding Is Challenging Gatekeeping in the Digital Age.” *Journalism Studies* 18 (9): 1087–1105.

Prior, Markus. 2005. “News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout.” *American Journal of Political Science* 49 (3): 577–592.

Prior, Markus. 2009. “The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure.” *Public Opinion Quarterly* 73 (1): 130–143.

Rai, Prerna, and Arvind Lal. 2016. “Google Pagerank Algorithm: Markov Chain Model and Hidden Markov Model.” *International Journal of Computer Applications* 138 (9): 9–13.

Sappelli, Maya, Dung Manh Chu, Bahadir Cambel, David Graus, and Philippe Bressers. 2018. “SMART Journalism: Personalizing, Summarizing, and Recommending Financial Economic News.” In The algorithmic personalization and news (APEN18) workshop at ICWSM18.

Scharkow, Michael, and Marko Bachl. 2017. “How Measurement Error in Content Analysis and Self-Reported Media Use Leads to Minimal Media Effect Findings in Linkage Analyses: A Simulation Study.” *Political Communication* 34 (3): 323–343.

Schmidt, Jan-Hinrik, Lisa Merten, Uwe Hasebrink, Isabelle Petrich, and Amelie Rolfs. 2019. “How Do Intermediaries Shape News-Related Media Repertoires and Practices? Findings from a Qualitative Study.” *International Journal of Communication* 13: 853–873.

Shoemaker, Pamela J., and Akiba A. Cohen. 2005. *News Around the World: Content, Practitioners, and the Public*. New York, NY: Routledge.

Tandoc, Edson C. 2014. “Journalism Is Twerking? How Web Analytics Is Changing the Process of Gatekeeping.” *New Media & Society* 16 (4): 559–575.

Taneja, Harsh, Angela Xiao Wu, and Stephanie Edgerly. 2018. “Rethinking the Generational Gap in Online News Use: An Infrastructural Perspective.” *New Media & Society* 20 (5): 1792–1812.

Tauchen, George. 1986. “Finite State Markov-Chain Approximations to Univariate and Vector Autoregressions.” *Economics Letters* 20 (2): 177–181.

Tewksbury, David. 2003. “What Do Americans Really Want to Know? Tracking the Behavior of News Readers on the Internet.” *Journal of Communication* 53 (4): 694–710.

Thorson, Kjerstin, and Chris Wells. 2016. “Curated Flows: A Framework for Mapping Media Exposure in the Digital Age.” *Communication Theory* 26 (3): 309–328.

Trilling, Damian. 2019. “Conceptualizing and Measuring News Exposure as Network of Users and News Items.” In *Measuring Media Use and Exposure: Recent Developments and Challenges*, edited by Christina Peter, Theresa Naab, and Rinaldo Kuhne, 297–317. Cologne: Halem.

Vermeer, Susan. 2018. “A Supervised Machine Learning Method to Classify Dutch-Language News Items.”

Wells, Chris, and Kjerstin Thorson. 2017. “Combining Big Data and Survey Techniques to Model Effects of Political Content Flows in Facebook.” *Social Science Computer Review* 35 (1): 33–52.

Wonneberger, Anke, Klaus Schoenbach, and Lex van Meurs. 2012. “Staying Tuned: TV News Audiences in the Netherlands 1988–2010.” *Journal of Broadcasting & Electronic Media* 56 (1): 55–74.

Zamith, Rodrigo, and Seth C. Lewis. 2015. “Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis.” *The ANNALS of the American Academy of Political and Social Science* 659 (1): 307–318.