UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Automated analysis of actor–topic networks on Twitter: New approaches to the analysis of socio-semantic networks

Hellsten, I.; Leydesdorff, L.

DOI
10.1002/asi.24207

Publication date
2020

Document Version
Final published version

Published in
Journal of the Association for Information Science and Technology

License
CC BY

Link to publication

Citation for published version (APA):
Hellsten, I., & Leydesdorff, L. (2020). Automated analysis of actor–topic networks on Twitter: New approaches to the analysis of socio-semantic networks. *Journal of the Association for Information Science and Technology*, *71*(1), 3-15. https://doi.org/10.1002/asi.24207

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Mar 2023

Automated Analysis of Actor–Topic Networks on Twitter: New Approaches to the Analysis of Socio-Semantic Networks

Iina Hellsten
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, PO Box 15793, Amsterdam, 1001 NG, The Netherlands. E-mail: i.r.hellsten@uva.nl

Loet Leydesdorff
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, PO Box 15793, Amsterdam, 1001 NG, The Netherlands. E-mail: l.a.leydesdorff@uva.nl

Social media data provide increasing opportunities for the automated analysis of large sets of textual documents. So far, automated tools have been developed either to account for the social networks among participants in the debates, or to analyze the content of these debates. Less attention has been paid to mapping co-occurrences of actors (participants) and topics (content) in online debates that can be considered as socio-semantic networks. We propose a new, automated approach that uses the whole matrix of co-addressed topics and actors for understanding and visualizing online debates. We show the advantages of the new approach with the analysis of two data sets: first, a large set of English-language Twitter messages at the Rio + 20 meeting, in June 2012 (72,077 tweets), and second, a smaller data set of Dutch-language Twitter messages on bird flu related to poultry farming in 2015–2017 (2,139 tweets). We discuss the theoretical, methodological, and substantive implications of our approach, also for the analysis of other social media data.

Introduction

Social media data provide social scientists with large textual corpora of complex social interactions in online debates. So far, quantitative methods and automated tools have been developed in two separate strands of network research. On the one side, in social network analysis the focus has been on networks of actors, and mapping the relations and structures of social interactions (Borgatti & Everett, 1997; Wasserman & Faust, 1994; Borgatti & Foster, 2003). On the other side, semantic network mapping has been used for analyzing the content of these messages. Content has been mapped in terms of patterns of co-occurring words (Danowski, 2012; Diesner, 2013), topics detected on the basis of clusters in word co-occurrence networks (for example, Carley & Kaufer, 1993; Courtial, 1994; Danowski, 2012; Diesner, 2013; Leydesdorff, 1989 and Leydesdorff, 1991), and implicit frames reflecting latent structures in word (co-)occurrences (Hellsten, Dawson, & Leydesdorff, 2010; Leydesdorff & Hellsten, 2005).

Both approaches—social network analysis and semantic network analysis—provide partial views of the communications in social media. Combining social and semantic networks can provide more comprehensive results for finding insights in online debates. The challenge of analyzing the co-occurrences of actors and topics in debates requires combining ideas from social and semantic network analysis. We propose an approach to mapping actor–topic networks using a "whole matrix," and discuss the relative merits of this approach in comparison to the 2-mode network-analysis approach of Borgatti and Everett (1997). Our approach is innovative both in terms of the network methods and its theoretical focus on mapping socio-semantic networks. The whole matrix approach enables us to map both heterogeneous and homogeneous sets of nodes and links in an integrated design.

First, in terms of methods, we improve on the 2-mode matrix approach as a representation of a bipartite network that is prominent in social network analysis (Everett &

Received November 22, 2017; revised October 3, 2018; accepted January 17, 2019

© 2019 The Authors. *Journal of the Association for Information Science and Technology* published by Wiley Periodicals, Inc. on behalf of ASIS&T. • Published online March 18, 2019 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.24207

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Borgatti, 2013). We propose to take into account the matrix of actors and topics attributed to tweets, and show the advantages of this approach in providing more informative results. Inspired by actor–network theory (Latour, 1996), we shift the focus from social actors and their semantics into co-addressing both actors and topics. Whereas social network analysis is interested in the interactions among authors of messages and the actors addressed by the authors, we focus on the interactions among the addressed actors and addressed topics, extracted from the contents of the messages. This shift in focus opens up new avenues for theory-building in the social sciences that is less focused on social actors as authoring messages, and more on addressing other actors in terms of topics.

Substantially, our focus is on Twitter messages, and we map the co-occurrences of hashtags (as representations of topics) and usernames (as addressed actors). Furthermore, we show an extension to a 3-mode approach that uses three different types of nodes (authors, addressees, and topics) in a single visualization. In summary, in addition to asking who (which author) used which concepts (topics), one can ask how actors and topics are co-addressed in communication. This research question builds upon earlier calls for combining actors and topics in actor–network theory (ANT), on the one hand, and semantic and socio-semantic network analysis, on the other.

Theoretical Framework: Network Approach

ANT was developed in the social-studies-of-science tradition from the early 1980s onwards, as a relational perspective on social interactions among both human and nonhuman agency. In the semiotic tradition, both semantics and social relations are considered as "actants" (Callon & Latour, 1981; Latour, 1996). Actants can represent human or nonhuman agents related in a network (Callon, 1986). In addition to the idea of both human and nonhuman actants, ANT, in a manner similar to social network analysis, theorizes networks using an encompassing relational and dynamic social theory.

Unlike social network analysis that focuses on interactions among human agents, ANT also focuses on nonhuman agents, and aims to "follow how a given element becomes strategic through the number of connections it commands, and how it loses its importance when losing its connections" (Latour, 1996, p. 372). Our approach focuses on the semiosis of connections in the social media debates instead of social relations among actors in the debates. We analyze usernames and hashtags addressed as actants in Twitter communications. In brief, we ask not who addressed which topics, but who was co-addressed with which topics. In the following we shall call the social agents originating communications "authors" and the actors addressed in the communications "addressees," while we refer to co-addressed topics and actors as "actants" following the actor–network terminology.

In order to position our approach in relation to the wider network theory, we first discuss two strands of network analysis. These two strands—social network analysis and semantic, co-word analysis—have been developed mainly at arm's length from each other (but see, Roth & Cointet, 2010; Roth, 2013). The challenge of theorizing meaningful socio-semantic networks and how they could change or enrich empirical research in the information sciences and communication studies has remained an open question.

In social network analysis, the methodology to measure interactions among social actors as "authors" has been elaborated over a number of decades (Wasserman & Faust, 1994). Bipartite networks of actors who are affiliated to social groups provide 2-mode affiliation networks of actors versus groups (Breiger, 1974). Computer programs make it possible to identify important authors in terms of their centrality in the networks. In social network analysis, social authors and their relations to each other have been studied, in addition to bipartite matrices of authors and their attributes (Borgatti & Everett, 1997). However, this methodology does not give access to the semantic content of the communications.

The content of communication has been the subject of semantic network analysis (Landauer, Foltz, & Laham, 1998) which has attracted growing scholarly attention since the early 1990s (Leydesdorff, 1989; Leydesdorff, 1991; Leydesdorff, 1997), in particular, in two distinct traditions—one thriving on human or computer-assisted coding, the second applying automated analyses to semantic co-word maps. Carley and Kaufer (1993), for example, called attention to combining the research fields focusing on symbols with semantic network analysis, arguing that these two representations were in need of crossfertilization. Later on, this approach was elaborated into systematic research on structures of concept networks using dedicated software packages (for example, AutoMap and ORA) that are based on the coding of words in the text(s) into categories including, for example, individual names, organization names, and other relevant categories (Diesner, 2013).

In particular, Diesner and Carley (2005) proposed the so-called meta-matrix approach to semantic network analysis. This approach and the related ORA software distinguishes among four content entities: (i) agents, (ii) knowledge categories, (iii) resources, and (iv) processes or tasks. The purpose of this design is to signal imbalances in the organizations. Technically, the meta-matrix approach combines affiliation matrices, while our approach focuses on the *de*composition of attribute matrices. In our opinion, the two approaches are analytically different and serve different objectives. Whereas the meta-matrix approach to semantic network analysis requires data cleaning, and manual or partly automated, vocabulary-assisted coding of the texts, our approach can be fully automated. After the coding, the meta-matrix approach can be used for automated network analysis of (large) sets of texts (Pfeffer & Carley, 2012). This approach, in our opinion, extends the range of manual and automated content analysis.

DOI: 10.1002/asi

In traditional manual content analysis (for example, Krippendorff, 1989), the focus is on explicit frames created ex ante by the coders when designing a coding scheme. Subsequently, the resulting networks of concepts (consisting of single words and/or phrases) represent the coders' interpretations of significant concepts instead of implicit or emerging meanings in the texts. In principle, such social-science-inspired text analysis is very similar to the quantitative methods developed in language studies such as cognitive linguistics (Sanders & Spooren, 2010).

Recently, automated analyses have been applied to both content analysis and semantic network analysis. Automated content analysis focuses on extracting associative frames of manually constructed actors and issues in documents (for example, Schultz, Kleinnijenhuis, Oegema, Utz, & van Atteveldt, 2012), and using automated cluster and sentiment analysis (for example, Burscher, Vliegenthart, & De Vreese, 2015). Factor analysis has been used for automated analysis of topics using a word/document matrix (Leydesdorff & Welbers, 2011; Vlieger & Leydesdorff, 2011). This factor-analytic approach is comparable to topic modeling that uses word distributions to detect topics, assigning words belonging to specific topics, and the co-occurrences of the words in topics, especially those using Latent Dirichlet Allocation (LDA), which assigns words into clusters using probability distributions (Blei, Ng, & Jordan, 2003). This method has been applied to the analysis of large sets of documents (for example, Jacobi, van Atteveldt, & Welbers, 2016).

In another strand of network semantics, Leydesdorff and Hellsten (2005) and (2006) developed automated semantic co-word maps to uncover the implicit frames in textual documents without human coding. This so-called vector-space model for mapping words is based on word/document matrices (Salton & McGill, 1983; Turney & Pantel, 2010). Using the word/document matrix, one takes into account not only dyads of co-occurring words, but also single words, triads, and so forth. In addition to the relations among co-occurring words, the method is able to take the positions of words in the vector space into account (for instance, see Leydesdorff & Hellsten, 2005). Nodes can occupy equivalent positions without entertaining a relation.

In addition to providing an application of ANT (Callon, Courtial, Turner, & Bauin, 1983), our approach provides an automated analysis of co-addressing actors and topics in text documents that can be widely applied to socio-semantic network analysis. We argue that topics and addressees can be represented as a 2-mode network of attributes instead of a bipartite network with two types of nodes, that is, in this case, a semantic network with two types of words (@usernames and #hashtags). In a next step, one can go beyond the ontology of ANT and consider addressees as potential authors of the Twitter messages, while hashtags are not able to "author" messages. In this respect, our ontology differs from ANT.

Whole-Matrix Approach

We operationalize the whole-matrix approach as follows. Each tweet can be considered as a unit of analysis to which both addressed actors (@usernames) and topics (#hashtags) are attributed. The resulting documents-versus-words matrix is asymmetrical, but one can generate an affiliations matrix of both hashtags and usernames in a single pass (by multiplication with the transposed of the matrix). The 2-mode matrix of hashtags versus usernames (as attributes) is contained in this matrix as off-diagonal subgraphs, whereas the co-hashtag and co-username matrices are positioned along the main diagonal (Figure 1 and Figure 2).

The matrix in Figure 1 is similar to a word/document matrix as used in library and information science (Salton & McGill, 1983) and also widely used in social network analysis (Borgatti & Everett, 1997) and recently also in semantic network analysis (for example, Yang & González-Bailón, 2017). Figure 2 shows the whole matrix containing the semantic network of actors and topics, and their relations in a single representation.

We argue that in the case of socio-semantic network analysis, the results of the whole matrix can provide more informative results than those based on the bipartite 2-mode matrix. In particular, the whole matrix approach enables us to capture both @username to #hashtag networks, and @mention to @mention or #hashtag to #hashtag networks, whereas the bipartite approach only captures @username to #hashtag networks. The off-diagonal subgraphs represent the intentions of the original authors to attach #hashtags to other

	hashtag 1	***hashtag 2***	***…***	***hashtag m***	***username 1***	***username 2***	***…***	***username k***
tweet 1								
tweet 2								
tweet 3								
tweet 4								
tweet 5								
….								
tweet n								

FIG. 1. "Hashtag/username" matrix of hashtags and usernames as attributes to tweets.

DOI: 10.1002/asi

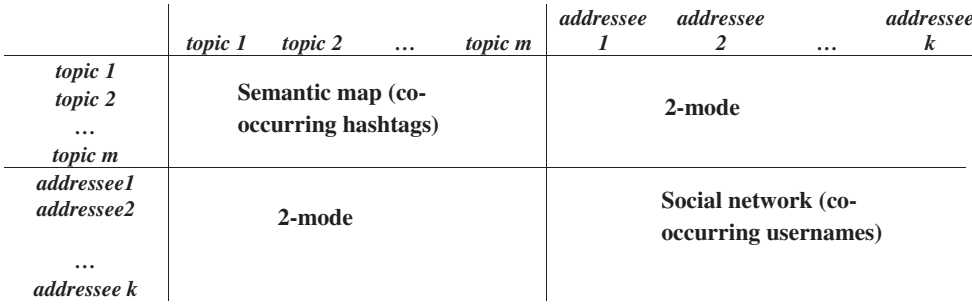

FIG. 2. Co-occurrence matrix of topics (hashtags) and addressed actors (usernames) as the whole-matrix approach.

tweet @mention users. We will demonstrate the surplus of this additional option and its possible extension to more than two dimensions in the Results section below.

Twitter Data

We chose to focus on Twitter data because Twitter provides users with the option to tag their tweets as belonging to specific topics by using #hashtags, and to address other users by @username. Hashtags can be used on Twitter to attach tweets into broader discussions and enable other Twitter users to follow specific topics and the related hashtags. (Bruns & Burgess, 2011; Bruns & Stiegelitz, 2013) We discuss the implications for using other types of data in the Discussion section.

In general, Twitter enables users to send short, maximally 140-character messages to other Twitter users—and recent upgrading allows for a maximum of 280 characters. The social media allows for addressing specific other users by adding the marker @ before the username of the targeted user; retweeting messages authored by other Twitter users, for example by using the mark RT at the beginning of the message; and for tagging messages using hashtags (with # mark) as well as spreading links to websites (using https://t.co/url). These Twitter-specific technological affordances (Foot & Schneider, 2006) allow for automated data extraction and subsequent analysis of the Twitter messages—and of the Twitter-specific functions. We discuss earlier findings related to the use of hashtags and usernames below.

Hashtags (for example, Bruns & Stieglitz, 2013; Perez-Altable, 2015; Holmberg & Hellsten, 2016) and hashtags in combination with keywords (boyd, Golder, & Lotan, 2010; Himelboim, Smith, Rainie, Shneiderman, & Espina, 2017) have been used for selecting a data set for analysis, and for identifying *ad hoc* publics on Twitter (Bruns & Burgess, 2011). boyd et al. (2010) showed that 36% of tweets contained a @username, but as few as 5% contained a #hashtag, whereas the more recent results by Gerlitz and Rieder (2013) presented 57,2% containing @usernames and 13% containing one or more hashtags.

In our data sets, both the average usage of @usernames and #hashtags is higher than in the earlier studies: whereas 88% of our Rio + 20 tweets contain a @username, in the bird flu data set 55% of the tweets address a @username. As regards #hashtags, tweets contain on average 1.,3 #hashtags in the Rio + 20 data set (130% of tweets containing a hashtag), and 1.1 #hashtags in the bird flu data set (110%). This indicates that both username and hashtag usage have increased over time. The increasing use of these Twitter-specific tools makes it important to automate the analysis of co-occurring hashtags and addressed usernames.

Saxton, Niyirora, Guo, and Waters (2015) manually coded the type of hashtags used by advocacy organizations and found that tweets containing hashtags used by several types of organizations were more likely to be retweeted. Less research has focused on how different types of institutional authors, such as nongovernmental organizations (NGOs) and political parties, use hashtags differently from individuals. Enli and Simonsen (2017) show that politicians use significantly larger numbers of hashtags in their tweets than journalists. Bruns and Steiglich (2013) show that hashtags are used more often in original tweets that are not retweets or replies to other users. Hashtags are also more often used in relation to major media events, such as royal weddings or the awarding of Oscars.

Earlier research has often focused on analyzing either co-occurring hashtags (for example, Russell et al., 2011; Gerlitz & Rieder, 2013) or co-occurring usernames in tweets (Ausserhofer & Maireder, 2013; Pearce, Holmberg, Hellsten, & Nerlich, 2014), but less on how these two co-occur in Twitter messages. On the use of usernames, Thelwall and Cugelman (2017) proposed a resonating topic method for evaluating the success of campaigns by the United Nations Development Programme (UNDP), and found that usernames are used in relation to mentioning others in the tweets as well as replying to other users, in particular in connection with the retweet symbol "RT@." We call both functions of using @usernames *addressing* other Twitter users. We included retweets in our data samples because retweets provide information on the amount of attention given to a particular issue.

In order to validate the approach, we apply the method to two data sets that differ in terms of (i) the size of the data set, (ii) the languages used in the tweets, and (iii) the types of discussion. Our large-scale data set consists of more than 100,000 tweets sent during the Rio + 20 meeting in Rio de Janeiro, Brazil, at the end of June 2012. This data set was collected using the open software crawler Webometric Analyst

DOI: 10.1002/asi

using the search term "#Rio + 20" (Thelwall, 2009).[1] Data thus collected can be opened in Excel and include a column for the language of each Twitter message. We used this language column to select all English-language Twitter messages for our analysis. Out of the total of 100,073 Twitter messages sent between 19 June and 2 July, 2012, 75,710 were in English. We further focus on the English-language tweets sent during the meeting between 20 and 22 June 2012. This resulted in a data set of 72,077 tweets that were further analyzed. Although the whole-matrix approach can be applied to virtually unlimited data sets, visualization of the resulting networks is restricted to roughly 100 nodes in order to keep the labels readable. In this sample of 72,077 tweets, 5,211 unique usernames and 3,150 unique hashtags were mentioned. In total, #hashtags were used 96,940 times in the sample of 72,077 tweets, whereas @usernames were used 63,475 times in the data set.

Our second data set of Twitter messages was collected using the software tool Coosto from the period of 1 June 2015 to 1 June 2017 using the search term "vogelgriep AND pluimvee" ("bird flu AND poultry"). The Coosto software tool requires the use of the Boolean search string to contain the word "en" ("and" in English) in the search. Unlike some other software tools, this does not mean that the results would have to contain the word "and." We downloaded 2,139 Twitter messages that include 234 unique @usernames and 230 unique #hashtags. The data set is in Dutch, but we discuss the results in English. In total, #hashtags were used 2,368 times, and @usernames 1,182 times in the data set of 2,139 tweets. For a more detailed analysis of a sample of 704 tweets using this method, see Hellsten, Jacobs and Wonneberger (2019).

Methods

We developed two dedicated computer programs—tweet.exe and frqtwt.exe—that are available at https://leydesdorff.github.io/twitter. Frqtwt.exe reads a file (named "text.txt") as input and provides a word frequency distribution. The analysis does not require the use of a stopword list for data cleaning since all the usernames and hashtags can be considered meaningful. Alphabetical ordering of the words results in #hashtags positioned at the top of the word frequency list, followed by @usernames. One can select the hashtags and the usernames to separate files for setting respective thresholds; that is, the smallest number of occurrences of the hashtags and usernames, if so wished.

Second, the routine tweet.exe reads the file "words.txt," which is compiled on the basis of the word frequency list, in combination with "text.txt," and generates the matrices shown in Figures 1 and 2. The resulting co-occurrence matrix of documents (tweets) versus words (hashtags and usernames) can be analyzed and visualized using software packages such as Pajek (for example, de Nooy et al., 2011) and VOSViewer (Van Eck & Waltman, 2011), respectively. We will compare the results of the bipartite 2-mode and the whole-matrix approaches using the Kamada and Kawai's (1989) algorithm as implemented in Pajek for the layout and VOSViewer for the visualizations.

Results

We discuss first the results using the small data set on bird flu and poultry in The Netherlands, and thereafter the results using the large data set of Twitter messages sent during the Rio + 20 environmental meeting in 2012. The United Nations conference on Sustainable Development, also called the Earth Summit and the Rio + 20 meeting, took place in Rio de Janeiro 20 years after the Rio meeting in 1992 that placed climate change on public and policy agendas as one of the main global threats. In both cases, we first discuss the similarities between the bipartite 2-mode and the whole-matrix approach, and thereafter highlight the differences between the two analyses. In the end, we will show a further application of the method that results in a 3-mode network of Twitter authors (usernames sending the messages) as an additional layer to the co-addressed hashtags and usernames in the Rio + 20 case.

Bird Flu Tweets

Bird flu epidemics have affected poultry farming, but also occasionally caused epidemics with human infections, most prominently in 2005–2006 when the H5N1 avian influenza virus spread from poultry to humans in Asia. Bird flu virus has infected poultry farms in Europe, causing poultry farms to keep their poultry inside as well as regulations to temporarily stop or restrict the import of chicken from infected areas and the transport of poultry. We focus on Twitter discussions concerning bird flu in poultry in The Netherlands during the period 2015–2017.

There were two peaks in the number of tweets during this period, in December 2015 related to new cases of the disease in poultry farms in France, and in November–December 2016, related to cases in The Netherlands (Hellsten, Jacobs, & Wonneberger, 2019). For pragmatic reasons, to limit The number of nodes in the resulting visualizations roughly to 100 nodes, we set the threshold to hashtags and usernames that appear five or more times in the data set. Using our dedicated software, however, the user is free to set this threshold lower or higher depending on a specific research question, the size of the data, or the purpose of the study. For example, one might be interested in the diversity of hashtags and take samples of specific hashtags and/or usernames, and compare then across case studies.

Both Figures 3 and 4 show the main hashtags #vogelgriep and #pluimvee located centrally in the network together with the main organization that is targeted in the tweets @pluimveeTweet. The latter is an online newsfeed designed

[1] We are grateful to Mike Thelwall for collecting the data set in 2012; only with this new method has it become possible to analyze the Rio + 20 tweets in a meaningful way.

DOI: 10.1002/asi

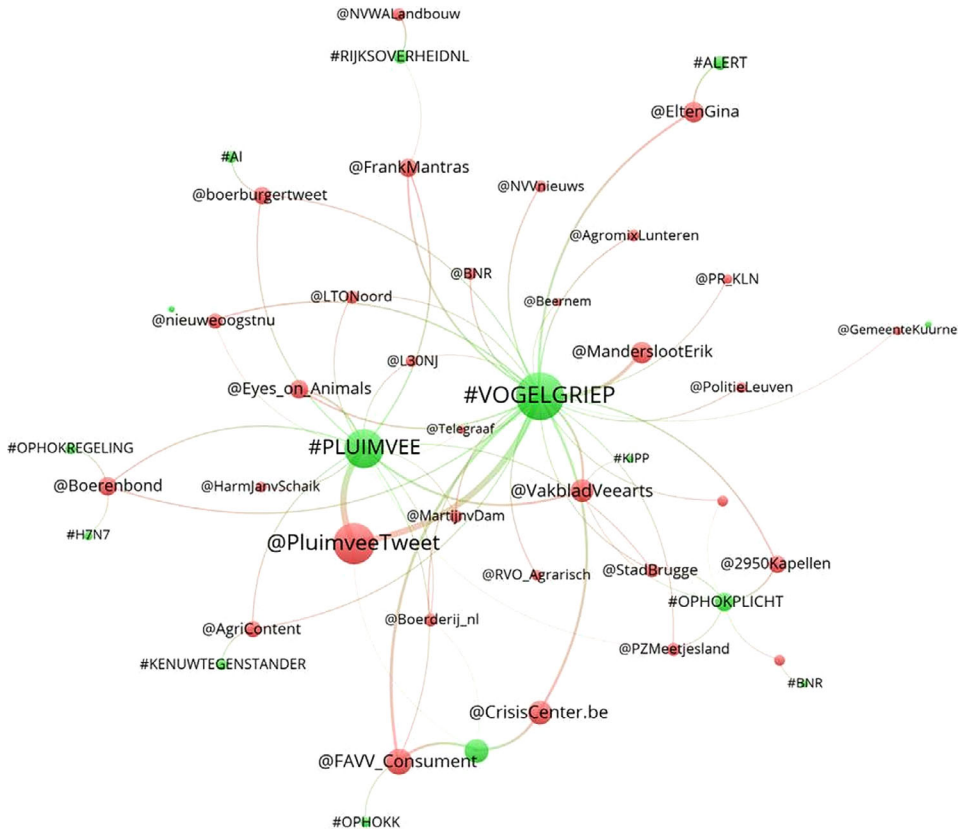

FIG. 3. Visualization on the basis of the bipartite 2-mode matrix of 39 hashtags (green) and 63 usernames of addressees (red) used ≥5 times in 2,139 Twitter messages on "bird flu and poultry"; the largest component, that excludes isolated nodes, contains 47 actants; VOSviewer was used for the layout. Node size represents the frequency of use of the word and line thickness the frequency of co-occurrences between the words. [Color figure can be viewed at wileyonlinelibrary.com]

for poultry farmers. It is noteworthy that both Figures 3 and 4 do not contain username-to-username connections. The network is highly dominated by the organizing, central hashtags. Updates of the situation were often retweeted in this data set; for example, in the tweet about new regulations that would be put in place the following day:

> RT @DNPPROVANT: Update vogelgriep: maatregel gaat in vanaf morgen @FAVV_Consument https://t.co/PVMhT9p6fX

In most tweets, both the hashtags vogelgriep ("bird flu") and pluimvee ("poultry") are used together so that the tweet is tagged for both issues. For example, the newsfeed PluimveeTweet was the first to send out the Twitter message on new cases of H5N1 epidemics in France, using both of the most common hashtags in the same tweet:

> Hoogpathogene H5N1 #vogelgriep vastgesteld in Frankrijk https://t.co/PdJzjwScqK #pluimvee

The map consists of a relatively large number of online news media (for example, @PluimveeTweet, "poultryTweet" @GriepTweets, "fluTweets" and @LandbouwNieuws, "agricultureNews") and municipalities (#Kapellen, #Deerlijk, #Heist-op-den-Berg, #Nijmegen) affected by the bird flu at poultry farms. Both figures also show the same clusters around FAVV_Consument that is affiliated with the Belgian Federal Agency for the Safety of Food, as well as the main regulations #ophokplicht and #ophokken ("indoor containment of the poultry").

However, the visualization of the bipartite network (Figure 3) loses these regional clusters of hashtags, such as Nijmegen, a Dutch city located in the province of Gelderland as connected to #NieuwsTwitter, another online newsfeed (separate cluster on the left-hand side), #griep ("flu"), and #nieuws ("news") in Figure 4. In other words, the bipartite 2-mode visualization cuts off clusters consisting of only a single type of node—hashtags in our case, and hence fails to map tweets such as:

> #Nijmegen Landelijke maatregelen vogelgriep alleen nog voor pluimvee, water- en loopvogels https://t.co/TaMs30tisW #nieuwstwitter

This tweet provides information about national regulations for poultry, waterfowl, and flightless birds in Nijmegen, tagging both the city of Nijmegen and one of the main newsfeeds, NieuwsTwitter. The bipartite 2-mode analysis (Figure 3) loses 20 actants when compared with the whole matrix (Figure 4).

The types of discussions (for example, crisis, a summit, long-term policy debate, and so forth) may result in different

DOI: 10.1002/asi

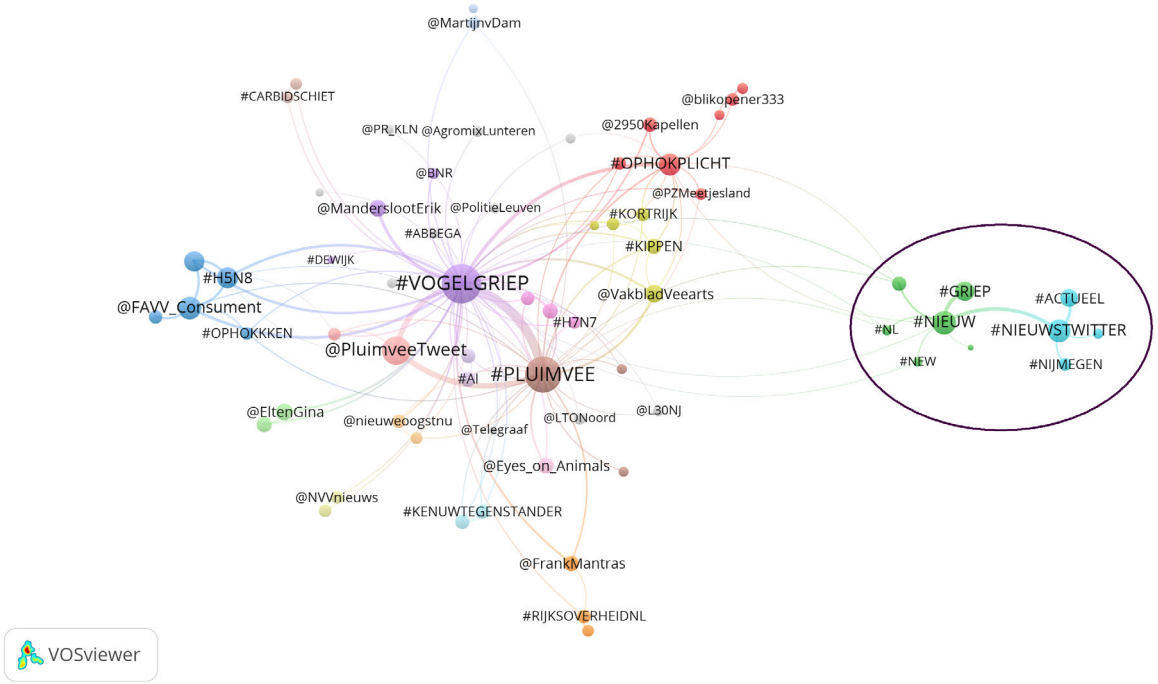

FIG. 4. Visualization based on the whole matrix of 39 hashtags and 63 usernames used ≥5 times in 2,139 Twitter messages on "bird flu and poultry"; largest component contains 67 actants; visualization: VOSviewer was used for the layout and clustering. Node size represents the frequency of use of the word and line thickness the frequency of co-occurrence between the words. [Color figure can be viewed at wileyonlinelibrary.com]

types of hashtag-username networks, since other types of actors can be prominent in other discussions. The map also shows NGOs active in environmental issues, such as Eyes_on_Animals, concerned with the effects of bird flu on food production. Such organizations are positioned on the periphery of the map due to their lesser role in the Twitter discussion on bird flu. see also Hellsten, Jacobs and Wonneberger (2019).

Our method provides an analytical tool to inspect how different types of actors are co-occurring with hashtags in addition to focusing on how specific authors use hashtags. The results can be used in crisis management to identify the national, regional, and local newsfeeds used by different organizations and citizens on Twitter for spreading information. In comparison, the whole-matrix approach also shows clusters of one type of node (for example, hashtags), while the bipartite 2-mode approach cuts these off from the main component. The whole-matrix approach informs us more completely than the network based on the bipartite 2-mode approach.

Rio + 20 Tweets

To further validate the method, we use a large data set of tweets sent during the United Nations Conference on Sustainable Development—the Rio + 20 meeting—that took place in 2012. This meeting is also called the Earth Summit or the RioPlus20 meeting, as it took place 20 years after the Rio 1992 meeting on biodiversity conservation and climate change. The tweets sent during the Rio + 20 meeting consist of a wide variety of participants discussing with one another during the meetings (for example, locations of lunch meetings, general reporting during the speeches, and about the meeting in general), the media sending out live information during the meeting, and political bodies trying to influence public opinion. This provides us with a large data set of more than 72,000 tweets during a short-term event that we would expect to consist of a high diversity of subtopics discussed. Since the data were collected with the search term #RioPlus20, all the tweets contain by definition this hashtag; we removed this hashtag from the analysis (see Figures 5 and 6).

In both the bipartite 2-mode and the whole-matrix visualization (Figures 5 and 6), one of the most prominent hashtags is #futurewewant; it is pronouncedly present in both visualizations. This hashtag connects several main actors during the meeting, such as @UN and @UNNewscenter. As an example, the hashtag has been used to retweet a message by WWF Australia and co-hashtagged with the general term @RioPlus20:

> RT @WWF_Australia: .@UN_Rioplus20 We want a game changing set of commitments that will ensure a future w food, water & energy for all @#futurewewant #RioPlus20

Both maps show several subtopics around energy issues (#energy, #energyforall, and @SGEnergyforall) and about women (#womenrio, @UNwomen). Global environmental NGOs, such as Oxfam, Greenpeace, and the World Wildlife Foundation (WWF) are present in both visualizations. The NGO Greenpeace has also been co-addressed with a major newspaper, @guardian.

DOI: 10.1002/asi

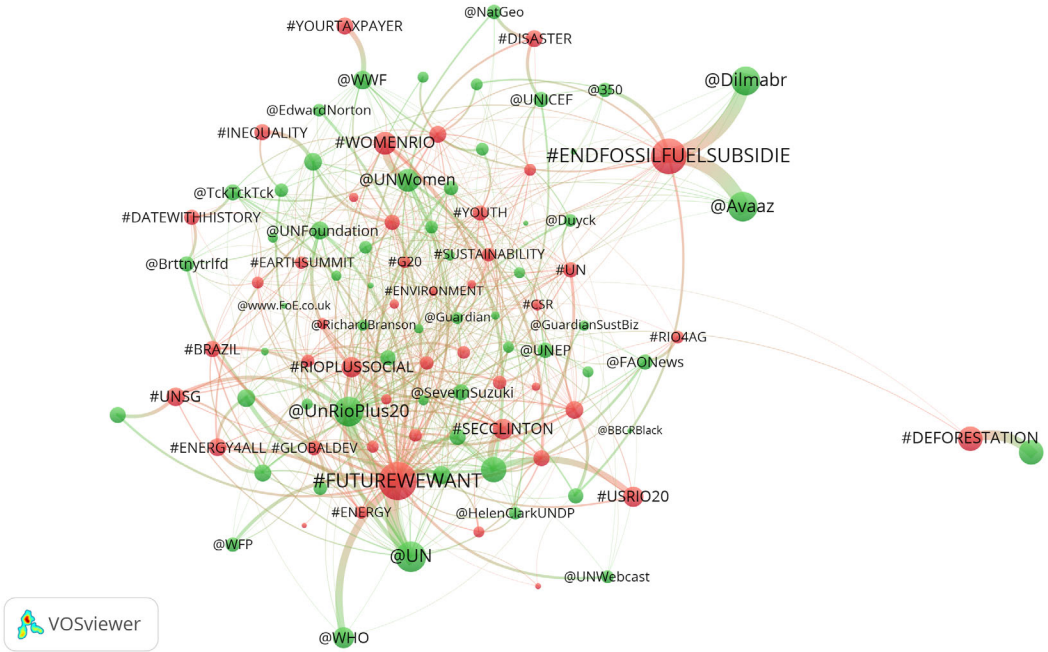

FIG. 5. Visualization of the bipartite 2-mode matrix of the 47 hashtags (red) and 58 usernames (green) used ≥150 times in the 72,077 English-language Twitter messages sent during the Rio + 20 meeting on 20–22 June 2012:largest component of 103 actants in the visualization; VOSviewer was used for the layout. Node size represents the frequency of use of the word and line thickness the frequency of co-occurrence between the words. [Color figure can be viewed at wileyonlinelibrary.com]

@Greenpeace moves to 'war footing' at #RioPlus20 http://t.co/nGjExgrN via @guardian

Both maps also show a strong activist cluster around the #endfossilfuelsubsidies linked to the actors @Avaaz and @dilmabr, the latter being the username of the former President of Brazil (on the right side in Figure 5, and on the left in Figure 6).

RT @Avaaz: You can find photos from our #EndFossilFuelSubsidies activities on Facebook: http://t.co/2qJ0Lcre & Flickr http://t.co/HXgNck4x #RioPlus20

However, the bipartite 2-mode matrix loses the connection between @Avaaz and @dilmabr in Figure 5. Similar to the bird flu and poultry case above, this is caused by omitting the

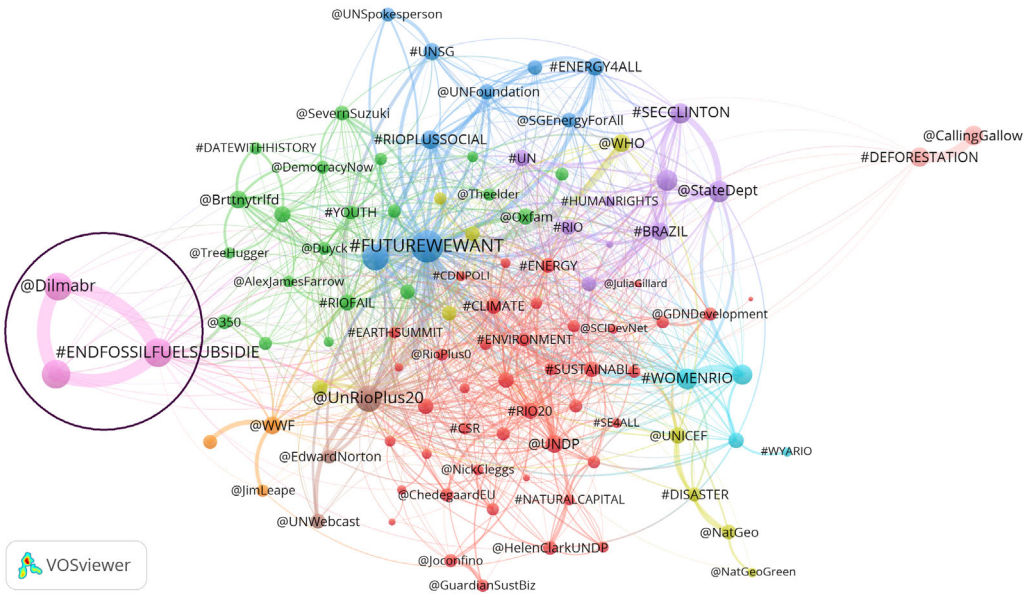

FIG. 6. Visualization based on the whole matrix of the 47 hashtags and 58 usernames used ≥150 times in the 72,077 English-language Twitter messages sent during the Rio + 20 meeting on 20–22 June 2012; largest component of 104 actants in the visualization; VosViewer was used for the layout and clustering. Node size represents the frequency of use of the word and line thickness the frequency of co-occurrence between the words. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

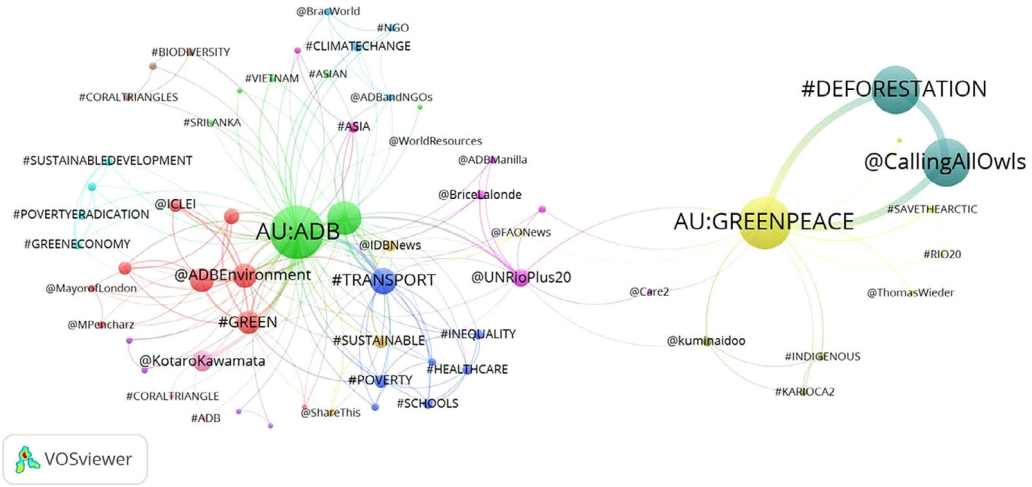

FIG. 7. Visualization on the basis of the 3-mode network of the two main organizations (Greenpeace and ADB) as "authors" and the hashtags and usernames addressed in their 173 and 160 tweets during the Rio + 20 meeting. Main component of 61 actants and two authors; VOSViewer was used for the layout and clustering. Node size represents the frequency of use of the word and line thickness the frequency of co-occurrence between the words. [Color figure can be viewed at wileyonlinelibrary.com]

connections among the same types of nodes; in the Rio + 20 case links between @usernames are not included.

Adding Authors to Hashtag-Username Networks

The analysis can be further elaborated, for example, by selecting tweet authors who have frequently posted on the issue, and then focusing on the co-occurring usernames and hashtags in the tweets by a specific active Twitter user, or organization, authoring Twitter messages (Hellsten, Jacobs & Wonneberger, 2019). This further refining is particularly useful in the case of large and heterogeneous data sets, such as the Twitter messages during an international meeting. As an example, we selected tweets that were sent out by two different types of organizations that authored more than 150 tweets during the 3-day meeting in Rio. One can add the authors as an additional (third) set of attributes to the right side of the whole matrix (Figure 1).

We selected Greenpeace, which authored in total 173 tweets during the conference (combined from its different Twitter username accounts, such as Greenpeace_de, Greenpeace_UPA, GreenpeaceCA, and GreenpeaceNZ), and the Asian Development Bank (ADB) which sent out 160 tweets in our data set (combined from the different local Twitter user accounts of the bank, such as ADB_Manila, ADBandNGOs, ADBClimate, and ADBEnvironment). The 173 tweets authored by Greenpeace during the 3-day meeting used 15 unique hashtags and 15 unique usernames twice or more times, whereas the 160 tweets authored by ADB make reference to 30 unique hashtags and 20 usernames used twice or more often. For both authors we included these hashtags and usernames addressed in the tweets with the prefix "AU:" (Figure 7).

Figure 7 shows that the two very active organizations (in terms of the number of tweets sent), Greenpeace and the ADB, mainly participated in their own subdebates during the meeting. The main shared hashtag is #futurewewant, which was also central in Figures 5 and 6. Both organizations also refer to shared usernames, such as @UNRioPlus20 and @FAONews.

Greenpeace was mainly co-addressing the topics of #RioPlus20 and #deforestration, linked with the username @CallingAllOwls that refers to a campaign of painting owls to save forests in order to promote zero deforestation by 2020. A typical tweet sent by Greenpeace is shown below:

> Greenpeace is @CallingAllOwls - pls RT and @ it to leaders #RioPlus20 + Zero #deforestation. One of 1000 voices: http://t.co/K9WiD5R0

Interestingly, the main hashtag addressed by Greenpeace—#deforestration—remained isolated in the context of all the tweets sent during the Rio + 20 meeting (Figure 6 on the left side), which indicates that the campaign was not highly retweeted by the other Twitter users during the meeting.

The ADB, in turn, was involved in several topical discussions, such as #poverty, #inequality, #healthcare (lower left-hand side), and #greeneconomy #sustainabledevelopment (right-hand side):

> Poor #transport exacerbates #poverty and #inequality, inhibiting access to #schools, #healthcare, markets & job opportunities. #rioplus20

The results provide a more detailed view of the activities of the selected organizations as authors participating in the debates on Twitter. One advantage of further labeling of the data according to the authors of the tweets is that different author types can be compared in greater detail; for example, due to the smaller size of the subgraphs, it is possible to include hashtags and usernames that were used twice or more often in the network visualization.

In summary, this method can be extended into 3-mode or even higher-order network analyses because it takes

DOI: 10.1002/asi

into account the whole matrix, as presented in Figure 1. This is an improvement compared with the bipartite 2-mode approach of Borgatti and Foster (2003). The whole-matrix approach outperforms socio-semantic network analysis, where the two types of nodes are co-addressed. The bipartite 2-mode approach includes only clusters consisting of similar types of nodes.

It should be noted that (in 2012) the mark @ was used not only in combination with a username to address another user but also to designate a location, simply replacing the word "at":

> RT @makower: Ted Turner @ UN Foundation dinner: "Clean coal: Bullshit." #rioplus20

One is able to differentiate between these two uses of the @ symbol in the whole-matrix approach by manually changing or removing the @ place usage that refers to location from the data set. (In the 2015–2017 data set the mark @ was used exclusively in combination with a username, as a conventional way to address other Twitter users. Perhaps this indicates changes in the use of social media tools over time.) However, more research is needed to analyze in detail how the use of other social media tools beyond Twitter has evolved over time. Such developments pose new challenges for social scientists interested in longitudinal studies of social media content. We discuss further implications of the whole-matrix approach in the Discussion and Conclusion section.

Discussion and Conclusion

We have proposed a new methodology for analyzing Twitter messages by focusing on the co-occurrences of Twitter-specific #hashtags and @usernames instead of the words used in the content of the Twitter messages. Our approach has the advantage of making it possible to map which users were addressed in connection with which topics. This approach helps to solve the problem of semantic networks that have been criticized for producing "bags-of-words" that remain vague in terms of meaningful interpretations. We have shown the advantages of the whole-matrix approach in providing more complete results than the bipartite 2-mode approach, in particular by also including clusters that consist of either hashtags or usernames. The bipartite 2-mode matrix tends to cut off such clusters. In addition, the whole-matrix approach allows for extending the analysis from two types of nodes into n-mode networks ($n > 2$). As an example, we extended the analysis to a 3-mode network of authors, actors, and hashtags, and mapped the results in a single visualization (Figure 7). Using ANT, the sending authors can also be considered as attributes of the tweets. This semiotic perspective adds opportunities for researchers to focus on multiple types of nodes depending on their research questions.

For theory-building, mapping hashtags and usernames instead of the words used in the message contents provides a more informative overview of the online discussions; co-occurrences of specific actors related to hashtags provides information on which actors were addressed in relation to which topics, hence advancing ANT by, indeed, analyzing hashtags and actors as "actants" based on their connections (Latour, 1996). In the context of ANT (Callon, 1986; Latour, 2005), these results are first steps toward automating the analysis of socio-semantic networks using text documents, in a way that does not rely on social networks between authors. Our approach makes visible the connections between actors and topics in online discussions. As our approach does not require focusing on the most active Twitter users, we are able to account for relations in which actors and topics ae addressed as co-occurring "actants." Further theory-building for the implications of our empirical research is needed.

To the emerging field of socio-semantic networks, previously applied to both offline (Saint-Charles & Mongeau, 2018; Basov, Lee, & Antoniuk, 2017) and online communications (Roth, 2013; Roth & Cointet, 2010), our approach offers a new empirical method for studying small as well as large-scale data sets in a way that provides meaningful results for the co-addressed actants in the communications. To our knowledge, this is the first automated effort to investigate how actors and topics are co-addressed in mediated communications.

Furthermore, our approach marks an improvement to the bipartite 2-mode approach that has been applied in social network analysis as the main methodological approach since the 1990s (Borgatti & Everett, 1997). Whereas this 2-mode approach has proven fruitful for the analysis of bipartite graphs, for example, of authors and words, the whole-matrix approach seems to perform more inclusively for analysis by combining actors with topics. There is a need for further theoretical and methodological research into comparing the two approaches with different types of data sets.

In practical terms, one of the additional advantages of this approach is that it can be used without data cleaning, such as removing from the analysis plural forms of words, the stemming of words, or using a stopword list to remove less meaningful words (for example, "the," "a," "an," "he," "she," "it," and so forth). All hashtags and usernames are meta-data, which are meaningful without any need for cleaning. Future studies could also compare semantic co-word networks with hashtag-username networks for a detailed comparison of the two approaches. The routines are also not limited by the size of the data set; in our case they were applicable to smaller data sets of a few thousand tweets and to a data set of more than one hundred thousand tweets. This allows for a more reliable bottom-up approach to social-media discussions.

In conclusion, this approach can be applied to a wide range of theoretical traditions in the communication sciences, such as research into issue arenas (Hellsten, Jacobs & Wonneberger, 2019) as well as stakeholder analysis by focusing on the co-mentioning of actors in news media, social media, and organizational media in general. Although we applied the method to the Twitter messages

DOI: 10.1002/asi

under study, the approach can also be applied to, for example, scientific publications where subject headings or keywords (meta-data) can be considered as #hashtags and actors cited in the texts as @usernames. One could extract a list of keywords assigned to scientific articles and a list of cited actors from the contents of academic publications, use these lists to construct the words.txt, and run the analysis in a way similar to the one presented in this article. One could also combine social network analysis of the relations between the authors of the tweets with those targeted in the tweets. As a further step, the approach could be used for analyzing other types of texts by visualizing, for example, the organization names addressed in newspaper articles, similar to @username in Twitter messages. Alternatively, the approach can be used for scientific texts using subject categories or keywords as #hashtags and mentioned actor names as @usernames.

More research is needed to further validate and improve the method, and to find optimal ways to apply it, including meta-data of textual content that are not tweets and do not include # and @ markers in the texts. This empirical research can feedback into theory-building in the information and communication sciences and signals a shift from author-based approaches to text-based approaches.

Acknowledgment

We thank the anonymous referees for their detailed comments.

References

Ausserhofer, J. & Maireder, A. (2013). National politics on Twitter: Structures and topics of networked public sphere. Information, Communication & Society, 16(3), 291–314.

Basov, N., Lee, J.S. & Antoniuk, A. (2017). Social networks and construction of culture: A socio-semantic analysis of art groups. In Complex Networks & Their Applications (vol. 963, pp. 785–796).

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Borgatti, S.P. & Everett, M.G. (1997). Network analysis of 2-mode data. Social Networks, 19(3), 243–269.

Borgatti, S.P. & Foster, P.C. (2003). The network paradigm in organizational research: A review and typology. Journal of Management, 29(6), 991–1013.

boyd, d., Golder, S. & Lotan, G.. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In 2010 43rd Hawaii International Conference on System Sciences, Kauai, HI.

Breiger, R.L. (1974). The duality of persons and groups. Social Forces, 53(2), 181–190.

Bruns, A. & Burgess, J.E. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011. Reykjavik: University of Iceland.

Bruns, A. & Stieglitz, S. (2013). Towards more systematic *Twitter* analysis: Metrics for tweeting activities. International Journal of Social Research Methodology, 16(2), 91–108.

Burscher, B., Vliegenthart, R., & De Vreese, C. (2015). Framing beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. Social Science Computer Review, 34(5), 530–545.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed.), Power, action and belief: A new sociology of knowledge. London: Routledge & Kegan Paul.

Callon, M., Courtial, J.-P., Turner, W.A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. Social Science Information, 22(2), 191–235.

Callon, M. & Latour, B. (1981). Unscrewing the big leviathan: How actors macro-structure reality and how sociologists help them to do so. In K.D. Knorr-Cetina & A.V. Cicourel (Eds.), Advances in social theory and methodology. Toward an integration of micro- and macro-sociologies (pp. 277–303). London: Routledge & Kegan Paul.

Carley, K.M. & Kaufer, D.S. (1993). Semantic connectivity: An approach for analyzing symbols in semantic networks. Communication Theory, 3 (3), 183–213.

Courtial, J.-P. (1994). Co-word analysis of scientometrics. Scientometrics, 31(3), 251–260.

Danowski, J.A. (2012). Analyzing change over time in organizations publics with a semantic network include list: An illustration with Facebook. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 954–959).

De Nooy, W., Mrvar, A., & Batagelj, V. (2011). Exploratory social network analysis with Pajek: Revised and expanded (2nd ed.). Cambridge, UK: Cambridge University Press.

Diesner, J. (2013). From texts to networks: Detecting and managing the impact of methodological choices for extracting network data from text data. Künstliche Intelligenz, 27(1), 75–78.

Diesner, J. & Carley, K.M. (2005). Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. In V.K. Narayanan & D.J. Armstrong (Eds.), Causal mapping for information systems and technology research (pp. 81–108). Harrisburg, PA: Idea Group Publishing.

Enli, G. & Simonsen, C.A. (2017). Social media logic meets professional norms: Twitter hashtags usage by journalists and politicians. Information, Communication & Society, 21, 1081–1096. https://doi.org/10.1080/1369118X.2017.1301515

Everett, M.G. & Borgatti, S.P. (2013). The dual-projection approach for two-mode networks. Social Networks, 35, 204–210.

Foot, A.K. & Schneider, S.M. (2006). Web campaigning. Cambridge, MA, and London, UK: MIT Press.

Gerlitz, C. & Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. *M/C Journal*, 16(2). ISSN 14412616. Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/view/620Rieder.

Hellsten, I., Dawson, J., & Leydesdorff, L. (2010). Implicit media frames; automated analysis of public debate on artificial sweeteners. Public Understanding of Science, 19(5), 590–608.

Hellsten, I., Jacobs, S., & Wonneberger, A. (2019). Active and passive stakeholders in issue arenas: A communication network approach to the bird flu debate on Twitter. Public Relations Review, 45(1), 35–48.

Himelboim, I., Smith, M.A., Rainie, L., Shneiderman, B., & Espina, C. (2017). Classifying Twitter topic-networks using social network analysis. Social Media +Society, 3(1), 1–13.

Holmberg, K. & Hellsten, I. (2016). Twitter campaigns around the fifth IPCC report: Campaign spreading, shared hashtags, and separate communities. SAGE Open, 6(3).

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, 4(1), 89–106.

Kamada, T. & Kawai, S. (1989). An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1), 7–15.

Krippendorff, K. (1989). Content analysis. In E. Barnouw, G. Gerbner, W. Schramm, T. L. Worth, & L. Gross (Eds.), International encyclopedia of communication (Vol. 1, pp. 403–407). New York, NY: Oxford University Press.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2), 259–284.

Latour, B. (1996). On actor-network theory: A few clarifications. Soziale Welt, 47(4), 369–381.

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. New York: Oxford University Press.
Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. Research Policy, 18(4), 209–223.
Leydesdorff, L. (1991). In search of epistemic networks. Social Studies of Science, 21, 75–110.
Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. Journal of the American Society for Information Science, 48(5), 418–427.
Leydesdorff, L. & Hellsten, I. (2005). Metaphors and diaphors in science communication: Mapping the case of stem-cell research. Science Communication, 27(1), 64–99.
Leydesdorff, L. & Hellsten, I. (2006). Measuring the meanings of words in contexts: An automated analysis of "Monarch butterflies," "Frankenfoods," and "stem cells". Scientometrics, 67(2), 231–258.
Leydesdorff, L. & Welbers, K. (2011). The semantic mapping of words and co-words in contexts. Journal of Informetrics, 5(3), 469–475.
Pearce, W., Holmberg, K., Hellsten, I., & Nerlich, B. (2014). Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC working group 1 report. PLoS One, 9(4), e94785.
Perez-Altable, L. (2015). The Arab Spring before the Arab Spring: A case study of digital activism in Tunisia. Global Media Journal (Arab Edition), 4(1–2), 19–32.
Pfeffer, J. & Carley, K.M. (2012). Rapid modeling and analyzing networks extracted from pre-structured news articles. Computational Mathematic Organization Theory, 18, 280–299.
Roth, C. (2013). Socio-semantic frameworks. Advances in Complex Systems, 16(4 &5), 1350013.
Roth, C. & Cointet, J.-P. (2010). Social and semantic co-evolution in knowledge networks. Social Networks, 32, 16–29.
Saint-Charles, J. & Mongeau, P. (2018). Social influence and discourse similarity networks in workgroups. Social Networks, 52, 228–237.
Salton, G. & McGill, M.J. (1983). Introduction to modern information retrieval. Auckland, NZ: McGraw-Hill.
Sanders, T. & Spooren, W. (2010). Discourse and text structure. In D. Geeraerst & H. Cuykens (Eds.), The Oxford handbook of cognitive linguistics. Oxford, UK: Oxford University Press.
Saxton, G.D., Niyirora, J.N., Guo, C., & Waters, R.D. (2015). #Advocating for change: The strategic use of hashtags in social media advocacy. Advances in Social Work, 16(1), 154–169.
Schultz, F., Kleinnijenhuis, J., Oegema, D., Utz, S., & van Atteveldt, W. (2012). Strategic framing in the BP crisis: A semantic network analysis of associative frames. Public Relations Review, 38(1), 97–107.
Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. San Rafael, CA: Morgan & Claypool (Synthesis Lectures on Information Concepts, Retrieval, and Services (Vol. 1, No. 1).
Thelwall, M. & Cugelman, B. (2017). Monitoring Twitter strategies to discover resonating topics: The case of the UNDP. El Profesional de la Información, 26(4), 649–661.
Van Eck, N.J. & Waltman, L. (2011). Text mining and visualization using VOSviewer. ISSI Newsletter, 7(3), 50–54.
Vlieger, E. & Leydesdorff, L. (2011). Content analysis and the measurement of meaning: The visualization of frames in collections of messages. Public Journal of Semiotics, 3(1), 28–50.
Wasserman, S. & Faust, K. (1994). Social network analysis: Methods and applications. New York: Cambridge University Press.
Yang, S. & González-Bailón, S. (2017). Semantic networks and applications in public opinion research. In J.N. Victor, A.H. Montgomery, & M. Lubell (Eds.), The Oxford handbook of political networks (pp. 327–353). New York, NY, USA. Oxford University Press.

Appendix

TWEET.exe (available at https://leydesdorff.github.io/twitter) generates a word-document occurrence matrix, a word co-occurrence matrix, and (if so wished) a normalized co-occurrence matrix from a set of lines (tweets) and a word list. The output files can be read into standard software (like SPSS, UCInet/Pajek, and so forth) for the statistical analysis and the visualization.

Input Files

The program needs two information, namely, (a) the name of the file "words.txt" that contains the words (as variables) in ASCII format, and (b) a file "text.txt" in which each line provides a textual unit of analysis (for example, a tweet). The number of lines is unlimited, but each line can at the maximum contain 4,000 characters. Each line has to be ended with a hard carriage return (CR + LF). Save the file as plain text with CR/LF in Word or in an ASCII editor such as Notepad.

The number of words (variables) is limited to 1,024; but keep in mind that most programs (for example, Excel) will not allow you to handle more than 256 variables in the follow-up. The words have to be on separate lines, which are ended with a hard character return and line feed. (Save in Word as plain text with CR/LF or use an ASCII editor (Notepad) for saving the file.)

- One can build a word frequency list with Frqtwt.Exe. This program reads <text.txt > and allows for the specification of a stopword list in <stopword.txt>. The results are provided as uppercase in the file <wrdfrq.txt.
- Stopword.txt contains 429 stopwords (available at http://www.lextek.com/manuals/onix/stopwords1.html). Both lists—the lists of words and stopwords—have to be available in the same folder as frqtwt.exe. The program checks the words in their *current* form (that is, without corrections for the plural). If stopword.txt is available, these words will not be included.
- Tweet.exe runs in a DOS-type Command Box under Windows. The program and the input files—text.txt and words.txt—have to be placed in the same folder. The output files are written into this directory as well. Please note that existing files from a previous run are overwritten by the program. Save output elsewhere if you wish to continue with the materials.

Output Files

The program produces three output files. Matrix.txt can be read into Excel and/or SPSS for further processing. Two files with the extension ".dat" are in DL-format (ASCII) and can be read into Pajek or UCInet for network analysis and visualization. Pajek is freely available at http://mrvar.fdv.uni-lj.si/pajek/.

a. *matrix.txt* contains an occurrence matrix of the words in the texts. The words are also the variable names in the SPSS syntax file labels.sps. One can read *matrix.txt* into SPSS using the text wizard and run *labels.sps* thereafter.

The matrix is asymmetrical: it contains the words as the variables and the tweets as the cases. In other words, each row represents a tweet in the sequential order of the text numbering, and each column represents a word in the sequential order of the word list. (One may wish to sort the word list alphabetically before the analysis.) The words are counted as frequencies with +1 for each occurrence.

DOI: 10.1002/asi

b. *coocc.dat* contains a co-occurrence matrix of the words from the same data. This matrix is symmetrical and it contains the words both as variables and as row labels. The main diagonal is set to zero. The number of co-occurrences is equal to the multiplication of occurrences in each of the texts. (The procedure is similar to the routine "affiliations" in UCInet, but the main diagonal is here set to zero in this matrix.) The file *coocc.dat* contains this information in the DL-format that can be read by Pajek or UCInet.

c. Optionally: *cosine.dat* contains a cosine-normalized co-occurrence matrix of the words in the same data. Normalization is based on the cosine between the variables conceptualized as vectors (Salton & McGill, 1983). (The procedure is similar to using the file matrix.txt as input to the routine Proximity in SPSS.) The file *cosine.dat* contains this information in the Pajek format. The size of the nodes is equal to the logarithm of the occurrences of the respective word; this feature can be turned on in Pajek. Tweet.exe can be stopped after running coocc.dbf and coocc.dat if one does not need the cosine values.

DOI: 10.1002/asi