A tool for tracking the propagation of words on Reddit

Tom Willaert, Paul Van Eecke, Jeroen Van Soest, Katrien Beuls

CCR 3 (1): 117–132
DOI: 10.5117/CCR2021.1.005.WILL

Abstract

The data-driven study of cultural information diffusion in online (social) media is currently an active area of research. The availability of data from the web thereby generates new opportunities to examine how words propagate through online media and communities, as well as how these diffusion patterns are intertwined with the materiality and culture of social media platforms. In support of such efforts, this paper introduces an online tool for tracking the consecutive occurrences of words across subreddits on Reddit between 2005 and 2017. By processing the full Pushshift.io Reddit comment archive for this period (*Baumgartner et al., 2020*), we are able to track the first occurrences of 76 million words, allowing us to visualize which subreddits subsequently adopt any of those words over time. We illustrate this approach by addressing the spread of terms referring to famous internet controversies, and the percolation of alt-right terminology. By making our instrument and the processed data publically available, we aim to facilitate a range of exploratory analyses in computational social science, the digital humanities, and related fields.

Keywords: Language propagation, media, reddit, digital methods

Background: cultural information transfer and online media

The data-driven study of how cultural information propagates through online (social) media is an active area of research, marked by a focus on the diffusion of 'viral' phenomena such as disinformation, conspiracy theories, hate speech, and internet 'memes'.

Strands of this line of research can be traced back to the work of Richard Dawkins, who actuated the scientific study of cultural information transfer by coining the term 'meme' as a designation for 'a unit of cultural transmission or a unit of imitation' (Dawkins, 1976/2016). Examples of memes thus include 'tunes, ideas, catch-phrases, clothes fashions, ways of making pots [and] of building arches' (idem.), or, as Dawkins has recently stated, 'anything that [...] spreads through the population in a cultural way' (Fazal, 2018). From the perspective of Dawkins' memetics, digital information networks such as email and social media generate opportunities to empirically corroborate theoretical assumptions and hypotheses about the propagation of cultural phenomena (Heylighen & Chielens, 2009; Gleick, 2011, ch. 11; Brewer 2016). Pioneering studies in the early 2000s for instance charted memetic phenomena by examining electronic chain letters (Bennet et al., 2003; Chielens, 2003; Goodenough & Dawkins, 2002).

However, important criticisms have been levelled at the biologically-based model of memetics and how it considers cultural 'units' passing through humans and infrastructures, as this perspective undermines the role of human agency or the transformative capacity of media (see for instance Sampson, 2011). Work on online information diffusion from the fields of media studies and digital methods, the framework in which the present paper inscribes itself, therefore adopts a wider perspective that also takes into account the culture and materiality of online media (Rogers, 2013). This has yielded insights into the origins, transformation and propagation of potentially harmful ideas, imagery and vernacular in online media environments (Tuters et al., 2018; Tuters and Hagen, 2019), including perspectives on the production and transmission of 'internet memes' (Shifman, 2013). The data-driven study of these cultural phenomena, and how they spread and are shaped through different media is not devoid of methodological and technical challenges, and scientific and societal breakthroughs in these budding fields hinge in no small part on novel instruments for (social media) data mining, analysis and visualization (Rogers, 2019). In support of empirical explorations of social media and their contents, the present paper thus contributes a tool for the study of language propagation on the social medium Reddit to the growing ecosystem of democratic instruments for digital methods and (social) media observation.

Goals

In order to meet the need for instruments to study online cultural information flows, we propose an online exploration tool for tracking the propagation of words across subreddits on reddit.com between December 2005 and April 2017 (N = 76,629,207 words). In this case, we are primarily interested to see by which communities particular terms and concepts are adopted over time. We thereby consider each subreddit on Reddit as a unique community of users, defined by specific sociological, cultural and linguistic factors. The tool thus supports the exploration of the propagation of words, including text-based memes, as well as some of the cultural and material aspects of Reddit itself (viz. its subreddits). The dataset underpinning the tool was constructed by processing the full Pushshift.io Reddit comment archive for the given period (Baumgartner et al., 2020).

Our tool's targeted units of analysis are single words (unigrams). For any given word that has occurred at least once on all of Reddit during this period, barring a set of 128 highly frequent words (see Section 4), we trace per month the subreddits on which the word was used for the very first time. This monthly aggregation of the data is motivated by the size of the dataset, its multi-year scope, and the long-term perspective this offers. We then visualize the subsequent occurrences of the word across different user communities as a graph (Fig. 1).

By defining the activity of mapping language propagation as registering for each monthly timestamp those subreddits on which a word appears for the first time, we impose a series of constraints upon our approach. First, we disregard whether or not a word remains in use on a particular subreddit after its first occurrence (this information, however, can be retrieved from a frequency chart in the tool's interface). Second, no 'top-down' measure is implemented for how frequently a word needs to occur on a particular subreddit for it to count as a significant occurrence. The dataset underlying the proposed tool contains the word frequencies for each subreddit, so users of the interface can specify a minimum word frequency to make visualizations easier to interpret. Third, no strict linguistic constraints are imposed by the approach, but search is limited to individual terms (i.e. 'Crooked Hillary' as a bigram cannot be retrieved, but the unigrams 'Crooked' and 'Hillary' can).

The tool is geared towards facilitating exploratory analyses in computational social science, digital humanities and related fields. An effort was

made to build an instrument with a user-friendly interface that allows easy interaction with the data. This tool should thus be situated in the design space of democratic tools for media monitoring for a broader audience (see Section 4 and Section 8 for design choices). In support of a range of applications, research questions and hypotheses, data have been pre-processed in such a way that the propagation of any word in the Reddit corpus can be tracked (see Section 3). The examples in this paper, however, specifically explore the propagation of ideologically-marked terminology, notably the spread of neologisms (new words) associated with internet controversies (e.g. 'fappening' and 'gamergate') and words indicative of right-wing slang (e.g. 'kek' and 'redpill').

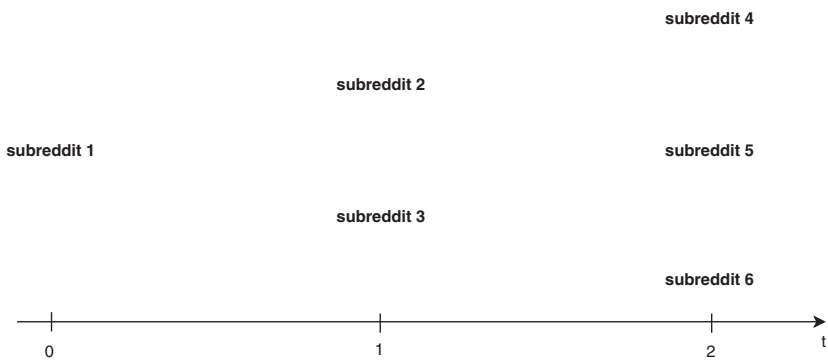

Fig 1. Schematic representation of the subsequent occurrences of a word across communities (subreddits) over time. At t = 0, the word first occurs on subreddit 1. At t = 1, the word occurs for the first time on subreddit 2 and subreddit 3. At t = 2, the word occurs for the first time on subreddits 4, 5, and 6.

Corpus data

While our data processing approach is suitable for a range of online media, the current implementation of the tool addresses language propagation on the social news site Reddit (reddit.com), also dubbed 'the front page of the Internet' (Lagorio-Chafkin, 2018). The motivation for selecting this particular corpus is threefold.

First, reddit.com is a well-documented site of cultural propagation that has been characterized as a 'meme hub' (Shifman, 2013, p. 13) and 'culture laboratory' (Lagorio-Chafkin, 2018). As of November 2017, the site was the fifth most visited site in the US, hosting content and discussions generated

by more than 430 million monthly active users in over 130,000 active communities (subreddits) (redditinc.com 2020, also see Amaya et al. 2019). This wealth of cultural, societal and linguistic data has been studied through a range of tailor-made methods and tools. This notably includes n-gram viewers for tracking the overall frequency of words and word groups over time (Olson, 2015; Redditor's Club, 2016; Olson and King, 2017), algorithms for latent semantic analysis (LSA) to map relations between (hateful) subreddits based on their linguistic properties (Martin 2016, 2017), and methods for creating taxonomies of trolls informed by typical word use (Squirrell, 2017). The proposed approach adds to these technical and methodological developments a means of studying the dynamics of word propagation, as well as the prominence (frequencies) of words down to the level of individual subreddits.

Secondly, Reddit has allowed users and user communities to create, maintain and moderate their own subreddits since this feature was introduced in 2008 (Ohanian, 2008). From a digital methods standpoint, this affordance of the platform allows for a combined perspective that deals with language propagation, as well as the material and cultural dynamics of Reddit itself, that is, it enables us to examine which subreddits enter the fold at specific times.

Thirdly, and crucially, recent data archiving and engineering efforts have opened up substantial collections of Reddit data for research. In the present case, the Pushshift Reddit dataset was used as a primary source for Reddit comments. Monthly compressed files containing dumps of all reddit comments from December 2005 up to and including April 2017 from the Pushshift.io online archive (Pushshift.io, 2020) were processed, resulting in a dataset of 12 years' worth of material from 3,169,506,937 Reddit comments (see 'Methodology and technical implementation' below).

Methodology and technical implementation

Construction steps

The construction of our tool comprised the following technical steps:

1) *Data gathering*: Reddit comments for the period between December 2005 and April 2017 were downloaded from the monthly archives in the Pushshift Reddit Database (Pushshift.io, 2020).
2) *Data cleaning and pre-processing:* For each month's worth of Reddit data, punctuation, URLs, and numbers were removed from the comment texts and case folding (removal of capital letters) was applied. The

monthly comments per subreddit were then tokenized. For each term in the dataset, we count for each month its frequency for each subreddit on which it occurs. The resulting data structure allows us to retrieve the sequence of subreddits in which the word occurs for the first time, along with the word's frequency on those subreddits.

3) *Data visualization:* The retrieved subreddit sequences are visualized as a scatter plot in which each subreddit is represented by a node. The node's position on the x-axis indicates the (monthly) timestamp, and nodes occurring at the same timestamp are evenly distributed among the y-axis for visibility reasons. Node sizes, finally, represent the word's frequency on the given subreddit for the given month. A dotted line connects monthly nodes for readability purposes. Note that the method holds no predictions about the origins and destinations of propagation, as well as the (causal) factors that might facilitate this spread.
4) *Web interface and analysis:* Visualizations can be rendered in an interactive web interface as part of the Penelope ecosystem of tools for computational social science (Penelope 2020, Willaert, Van Eecke, Beuls, & Steels, 2020). As illustrated in Fig. 2, the tool allows users to submit a keyword or list of keywords, the propagation graphs of which are then shown one below the other. This configuration allows for visual comparison of different queries. When a user clicks a node in the graph, the first post containing the keyword is retrieved from the pushshift.io API. This propagation visualization is complemented with a stream graph that shows the absolute frequencies that word over time.

Software information and pipeline diagram

The current implementation of the tool is available at https://penelope.vub.be/language-propagation/. It comprises a graphical web interface as well as a database endpoint that provides access to a MongoDB with the Reddit word count dataset (See Fig. 3). The interface that allows users to interact with the dataset is characterized by three design choices. First, users are given control over what they consider the minimum frequency that a keyword needs to have on a subreddit for it to be displayed as a significant first occurrence, as no such assumptions are implicit in the dataset. Second, we opt for a graphic representation reminiscent of a timeline in order to clearly indicate the sequence of first occurrences, complemented with a frequency flowchart. Finally, and importantly, when a user clicks on one of the subreddit names, the tool sends a request to the Pushshift.io API, which returns the full text of the first comment containing the keyword on that particular subreddit. This can give users an initial sense of the contextual

meaning of the term on the subreddit, as well as reveal potential shifts in that meaning over time.

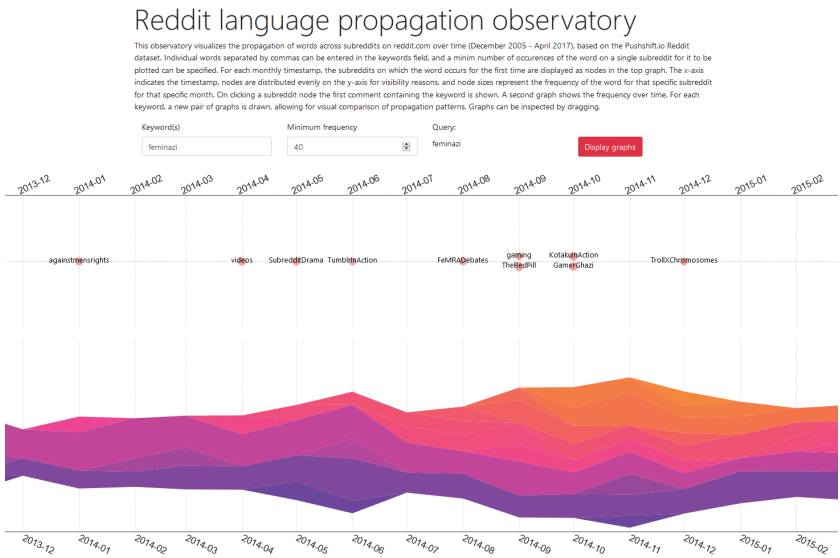

Fig. 2 Screenshot of the tool's web interface with graphs for the propagation (top graph with subreddit nodes) and frequency (bottom stream graph) of the slang term 'feminazi' over time. Within the timeframe shown, the term is seen to percolate from subreddits on men's rights through subreddits dealing with gaming, and the term's frequency peaks around November 2014 and April 2015. Only subreddits where the word has a minimum frequency of 40 are shown.

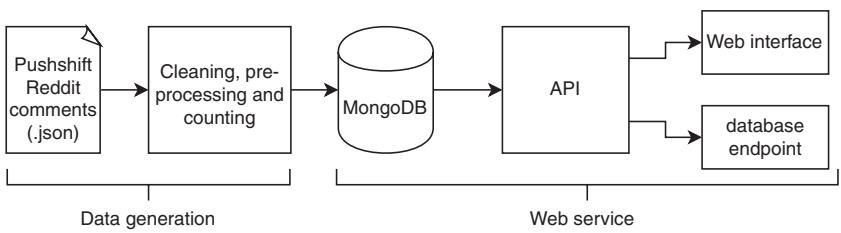

Fig 3. Technical pipeline diagram.

List of highly frequent words omitted from the corpus

For efficiency purposes, a set of 128 highly frequent words has been omitted from the corpus data:

a, about, after, all, also, am, an, and, any, are , as, at, back, be, because, been, best, but, by, can, could, d, deleted, did, do, don, even, find, first, for, from, get, go, going, good, got, great, gt, had, has, have, he, here, how, i, if, in , into, is, it, its, just, know, like, ll, look, love, m, make, me, more, most, much, my, need, new, no, not, now, of, on, one, only, or, other, out, over, people, post, pretty, r, re, really, right, s, see, should, so, some, something, still, sure, t, than , thanks, that, the, their, them, then, there, they, think, this, though, time, to, too, up, ve, very, want, was, way, we, well, were, what, when, where, which, who, will, with, work, would, you, your

Examples

The proposed tool for the tracking word propagation on Reddit can support a range of exploratory investigations in fields across the social sciences and humanities. This includes theory-building and hypothesis-testing in the domains of memetics and cultural information propagation, as well as linguistics. Similarly, exploratory overviews of how certain vernacular terms spread across subreddits can underpin media-historical narratives and sociological accounts of online communication. While in-depth analyses are beyond the scope of this methodological paper, the present section combines a selection of these perspectives in two examples aimed at illustrating potential uses and applications of the tool under discussion. Both examples were drawn from Christine Lagorio-Chafkin's extensive cultural history of Reddit, *We are the nerds. The birth and tumultuous life of reddit, the internet's culture laboratory* (Lagorio-Chafkin, 2018). The first example thereby explores the diffusion of novel terminology referring to a number of online controversies that have marked Reddit's history, such as 'the fappening' and 'gamergate'. The second example addresses the spread of vernacular terms associated with the alt-right. This includes examples such as 'kek', and 'redpill'.

Example 1: controversies

Reddit's history has seen a number of controversies characterized by the spread of contentious content, including explicit media, rumours or hate speech. An example of the former is 'the fappening', a term that users coined for the cascades of links to private nude photos stolen from the hacked iPhones of actresses, models, athletes and reality stars that were originally posted on 4chan, and emerged on Reddit in late August 2014 (with victims including Kate Upton, Kim Kardashian, Eva Longoria,

Jennifer Lawrence, and Cara Delevigne) (Lagorio-Chafkin, 2018, part IV, Ch. 5). As documented by Lagorio-Chafkin, the term 'fappening' '[…] stems from *fap*, for a sound effect indicating male masturbation, inspired by anime comics; the suffix *-penning* is a play on a popular hacker meme. "It's happening"' (idem.).

Using the tool under discussion, we can visualize the sequence in which communities started discussing the controversy by using the word 'fappening' (Fig. 4). This visualization confirms the peak of the controversy in September 2014, along with its subsequent spread. In addition to confirming

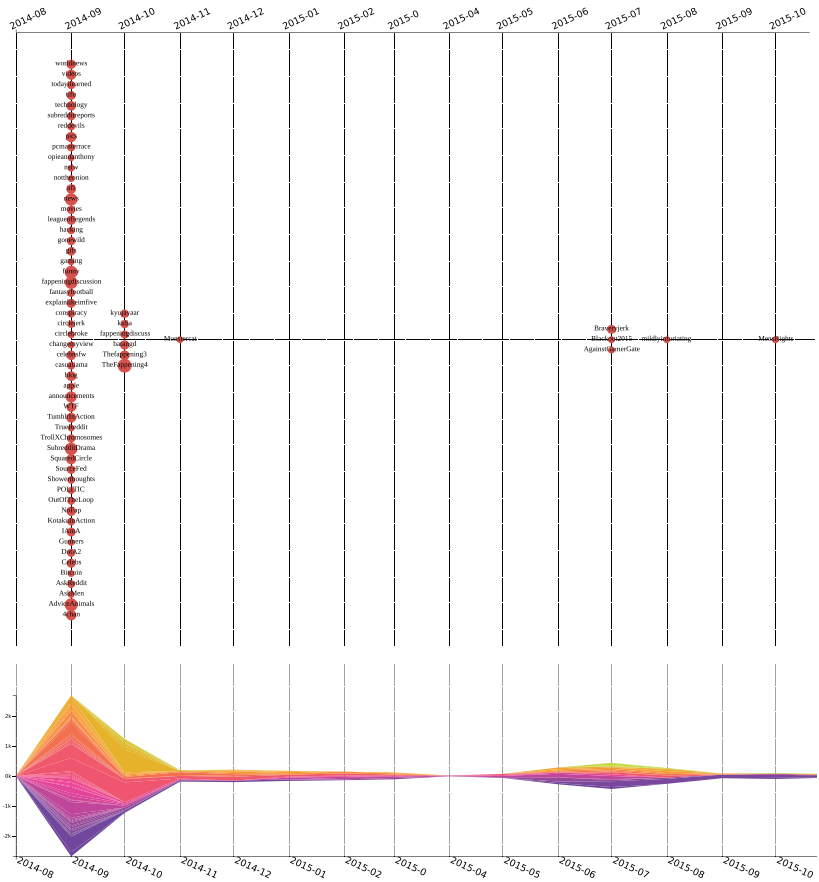

Fig 4. Visualization of the propagation sequence and streamgraph of the term 'fappening' (with minimum frequency set to 25). The term propagates to a large and diverse number of subreddits in September 2014, with the propagation to new subreddits diminishing again in October 2014.

the viral nature of the controversy, the visualization displays the names of the subreddits involved, allowing one to qualitatively assess the types of communities through which they have spread. With a minimum word frequency set to 25, it can for instance be observed that the controversy immediately propagates through a diverse range of subreddits, including communities dedicated to general news (r/worldnews, r/news), technology (r/apple, r/hacking, r/bitcoin), gaming (r/leagueoflegends), right-wing subreddits (r/4chan, r/AskMen), conspiracies (r/conspiracy), and groups dedicated to NSFW content (r/nsfw, r/gonewild). After the initial peak, the debate spreads to a series of subreddits dedicated to furthering the debate and preserving the leaked content (r/thefappening3, r/TheFappening4, r/fappeningdiscuss).

A second example of a controversy that involved Reddit around the same time as 'the fappening' was 'gamergate': a hate campaign against female game developers and critics (Lagorio-Chafkin, 2018, Part IV, Ch.6). Visualization of the propagation sequence of the term 'gamergate' reveals a 'viral' pattern similar to that of 'fappening' (Fig. 5). Closer inspection of the involved subreddits and communities where the word has a minimum frequency of 25 shows how the term initially figures on communities dedicated to tech and gaming (e.g. r/technology, r/truegaming, r/pcmasterrace, r/pcgaming), the alt-right (r/MensRights, r/4chan, r/Askmen), and conspiracies (r/conspiratard, r/conspiracy). In subsequent months, a further propagation can be observed to communities dedicated to minorities (e.g. r/gaymers, r/transgamers, r/AskFeminists, r/AskWomen), as well as the alt-right (r/TheRedPill, r/TheBluePill), in order to also enter more mainstream subreddits in subsequent months (e.g. r/ukpolitics, r/Europe, r/comicbooks, r/movies).

Example 2: right-wing vernacular

The internet controversies discussed earlier can be associated with the incubator for far-right, antagonistic slang that has formed on Reddit (see Lagorio-Chafkin, 2018, part V, Ch.7). The tool under discussion allows us to explore the different types of communities that have used these slang terms over time, which can be indicative of the words' continuously shifting meanings. We illustrate this by exploring the percolation of the words 'kek' and 'redpill'.

In the context of right-wing online discourse, the word 'kek' has been assigned multiple meanings. Originally borrowed from the game *World of Warcraft* where 'kek' figured as an in-game translation of 'LOL' ('laughing out loud'), the word was soon associated with the frog-like Egyptian god

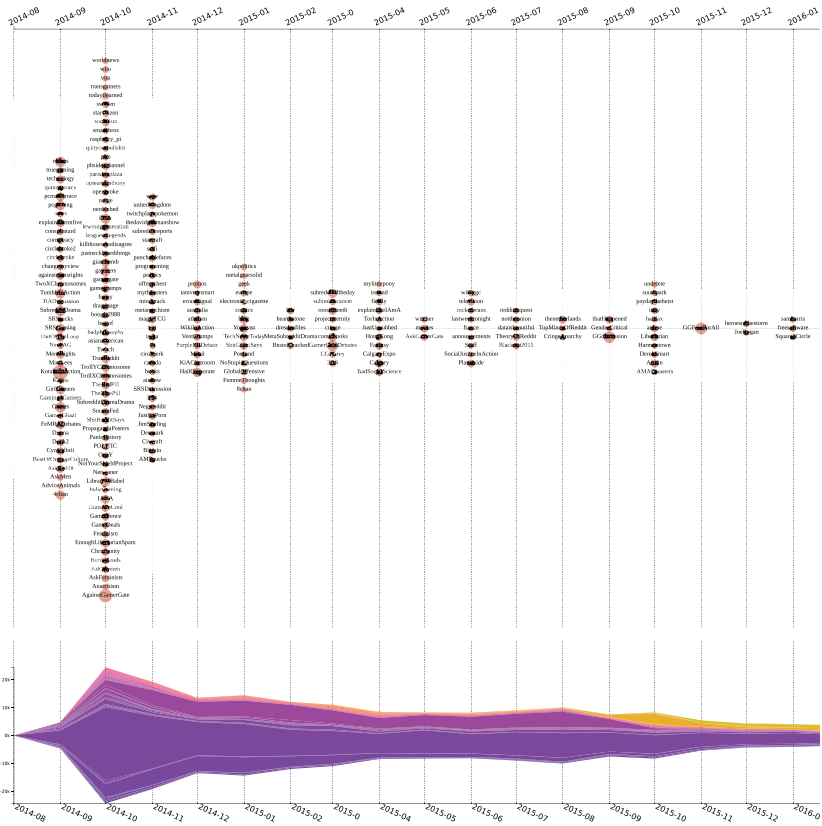

Fig 5. Visualization of the propagation sequence and streamgraph of the term 'gamergate' (with minimum frequency set to 25). Similar to the example of 'fappening', a rapid 'viral' spread is initially observed, peaking in October 2014, with the spread stabilizing during subsequent months.

of darkness 'Kek'. Through this connection with the figure of the frog, the term 'kek' then became associated with the meme of Pepe the frog as part of a mock religion (e.g. 'Praise Kek!') (knowyourmeme.com, 2020). An exploration of the subreddits where the word has a minimum frequency of 25 reveals how the term propagates from communities dedicated mainly to games and memes (e.g. r/twitchplayspokemon, r/starcraft, r/hearthstone, r/pokemon), and subsequently enters more politically-oriented subreddits (e.g. r/ukpolitics, r/politics, r/SandersForPresident), some of which explicitly right-wing (e.g. r/The_Donald) (Fig. 6).

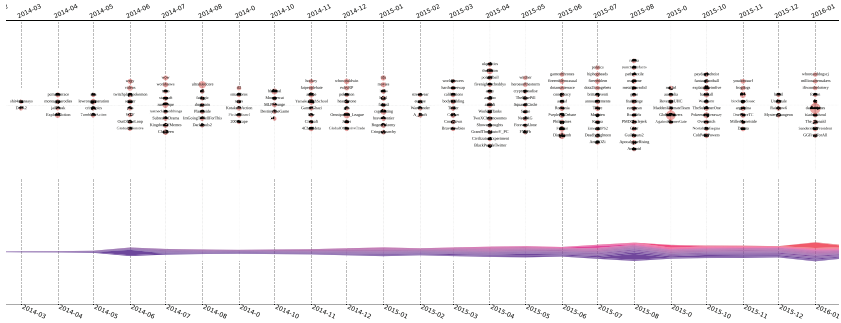

Fig 6. Segment of the propagation timeline and streamgraph showing 'kek"s spread from gaming communities (e.g. r/runescape and r/starcraft) into the political arena (e.g. r/The_Donald). Only subreddits where the term's minimum monthly frequency is 25 are shown.

The term 'redpill' originates from the famous 1999 movie *The Matrix,* where the protagonist Neo takes a red pill (as opposed to the blue pill) in order to see the world for what it really is. As elaborately documented in Hagen et al., 2020, the term was first integrated into the vernacular of the Men's Rights movement, and it was subsequently embraced on Reddit as a right-wing alternative to being 'woke', that is, seeing the truth of white nationalism. Throughout its history, the term has thus been appropriated to a range of different uses and contexts. The present tool for instance allows us to zoom in on the term's propagation ca. 2013-2014, confirming the observation that the term 'has origins in subreddits concerned with "men's-rights activism"(MRA) before moving to more explicitly politically oriented subreddits." (idem.). A minimum word frequency of 25 shows that following occurrences on the subreddit r/TheRedPill in the beginning of 2013, the term subsequently appears on such subreddits as r/AskMen, r/MensRights, r/againstmensrights, and r/everymanshouldknow (Fig. 7). Of these subreddits that – as their names suggest – cater to male communities, r/MensRights is most deliberately associated with MRA, r/againstmensrights is a 'watchdog' for the former, and the remaining subreddits can be characterized as more general advice subreddits. This suggests that the term actually quite rapidly propagated to more 'mainstream' subreddits.

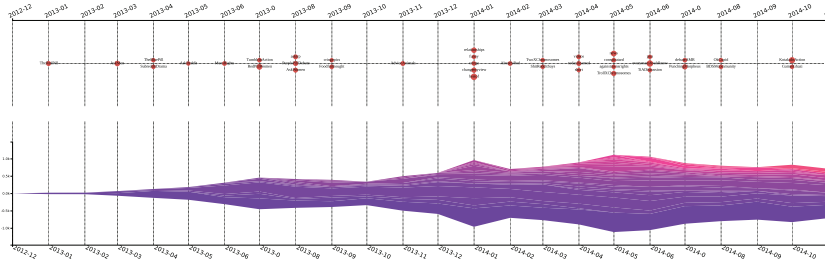

Fig 7. Early propagation timeline and streamgraph of 'redpill'. The visualization shows a consistent spread, and shows how the term initially spreads via 'men's rights' subreddits as well as more mainstream subreddits. Only subreddits where the term's minimum monthly frequency is 25 are shown.

Conclusions and avenues for future research

This paper has presented an online tool for visualizing the consecutive occurrences of words across communities on Reddit. Through showcases investigating the spread of online controversies and alt-right conspiracies, it has been demonstrated how this approach can support explorative research across disciplines. Given the exploratory objectives of the proposed tool and some of the restrictions this entails (including the monthly aggregation and the current implementation's focus on the period 2005-2017, which leaves some more recent neologisms out of scope), a number of pathways can be identified towards future, more in-depth studies of word propagation *patterns*.

First, it should be noted that the proposed tool is oriented towards observing sequences rather than *explaining* propagation patterns or revealing causal relations. Such explanations might however be supported by new media-theoretical insights, and by introducing additional context, notably by applying similar methods to other (social) media data (e.g. twitter, 4chan and 8chan).

Secondly, while the present tool allows for a visual comparison of subreddit sequences, deeper insights can be expected from deploying more quantitative descriptions of the patterns underlying these sequences. To this end, interesting parallels might for instance be drawn with the typology for algorithmic ranking patterns for Youtube videos on sociocultural issues proposed by Rieder et al. (2018). This typology distinguishes between 'newsy' patterns (with constantly changing rankings over time), 'stable' patterns (with a similar ranking of videos throughout the period) and 'mixed'

patterns (which consist of stable periods interrupted by 'newsy' patterns). To a certain extent, this typology also maps onto our respective examples on 'fappening' (newsy), 'kek' (stable), and 'gamergate' and 'redpill' (mixed). Furthermore, algorithms from evolutionary biology and related fields might yield more precise metrics for not only documenting the successful spread of concepts, but also registering which factors have contributed to this propagation (Shifman, 2013, p. 174-175; Adamic et al., 2014). Similarly, methods are needed to efficiently measure the similarity of propagation patterns, as this might reveal words or memes that propagate as a more complex group called a 'memeplex' (Shifman, 2013, p. 10).

Finally, the study of online language propagation and memetics can benefit greatly from empirical methods for detecting neologisms – that is, of methods that move beyond the mere detection of new word forms (such as those discussed above), but also take into account the different meanings that might be assigned to (existing) word forms in different context. To this end, the use of formalizations of word meaning, such as word embeddings, have already yielded interesting preliminary results for the study of online vernacular (Van Soest, 2019).

References

Adamic, L., Lento, T., & Ng, P. (2014). The evolution of memes on facebook. *Facebook Data Science*. Retrieved from https://www.facebook.com/notes/facebook-data-science/the-evolution-of-memes-on-facebook/10151988334203859/

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2019). New data sources in social science research. Things to know before working with reddit data. *Social science computer review*. https://doi.org/10.1177/0894439319893305

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift reddit dataset. arXiv:2001.08435

Bennet, C., Li, M., Ma, B. (2003). Chain letters & evolutionary histories. *Scientific American*, *288*(6), 76–81.

Brewer, J. (2016). A forty year update on meme theory. Retrieved from https://evolution-institute.org/blog/a-forty-year-update-on-meme-theory/

Chielens, K. (2003). *The viral aspects of language. A quantitative research of memetic selection criteria*. [Unpublished master's thesis], Vrije Universiteit Brussel. Retrieved from http://memetics.chielens.net/master/thesis.pdf

Dawkins, R. (2016). *The selfish gene. 40th anniversary edition. Oxford landmark science*. Oxford, United Kingdom: Oxford University Press.

Fazal, M. (2018). Richard Dawkins told us what he thinks about memes. *Vice*. Retrieved from https://www.vice.com/en_us/article/d35ana/richard-dawkins-told-us-what-he-thinks-about-memes

Gleick, J. (2011). *The information. A history, a theory, a flood.* New York, United States: Vintage Books.

Goodenough, O.R. & Dawkins, R. (2002). The 'St Jude' mind virus. *Nature* 371, 23–24.

Hagen, S., Tuters, M., & Wilson, J. (2020). Reactionary wokeness. How redpilling became a thing on Reddit. Retrieved from http://oilab.eu/reactionary-wokeness-how-redpilling-became-a-thing-on-reddit/.

Heylighen F. & Chielens K. (2009). Evolution of culture, memetics. In Meyers R. (Ed.), *Encyclopedia of Complexity and Systems Science* (pp. 3205-3220). New York, United States: Springer.

Knowyourmeme.com. (2020). Kek. Retrieved from https://knowyourmeme.com/memes/kek.

Lagorio-Chafkin, C. (2018). *We are the nerds. The birth and tumultuous life of Reddit, the internet's culture laboratory*. New York and Boston, United States: Hachette Books.

Martin, T. (2016). Interactive map of reddit and subreddit similarity calculator. Retrieved from https://www.shorttails.io/interactive-map-of-reddit-and-subreddit-similarity-calculator/

Martin, T. (2017). Dissecting Trump's most rabid online following. Retrieved from https://fivethirtyeight.com/features/dissecting-trumps-most-rabid-online-following/

Ohanian, A. (2008). blog.reddit – what's new on reddit: make your own reddit. Retrieved from https://web.archive.org/web/20140521025313/https://redditblog.com/2008/03/make-your-own-reddit.html.

Olson, R. (2015). The reddit ngram viewer. Retrieved from http://www.randalolson.com/wp-content/uploads/The_Reddit_Ngram_Viewer.pdf

Olson, R. & King, R. (2017). How the Internet Talks. Retrieved from https://projects.fivethirtyeight.com/reddit-ngram/?keyword=dank%20meme.rage%20comic&start=20071015&end=20150831&smoothing=10

Penelope (2020). The Penelope platform. Tools and techniques for computational social science. Retrieved from https://penelope.vub.be/

Pushshift.io. (2020). Directory contents (Reddit comment data). Retrieved from https://files.pushshift.io/reddit/comments/r/NoBullshitGaming. (2020). Game talk and reviews from regular people who like games. Retrieved from https://www.reddit.com/r/NoBullshitGaming/

r/NoBullshitGaming (2020). Game talk and reviews from regular people who like games. Retrieved from https://www.reddit.com/r/NoBullshitGaming/

Redditinc.com. (2020). Reddit by the numbers. Retrieved from https://www.redditinc.com/press

Redditor's Club. (2016). Redditor's club. Retrieved from https://github.com/yuguang/reddit-comments

Rieder, B., Matamoros-Fernández, A., Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures'. Investigating the modulation of visibility in YouTube search results. *Convergence. The International Journal of Research into New Media Technologies*. doi: 10.1177/1354856517736982.

Rogers, R. (2013). *Digital methods*. Cambridge, MA and London: MIT press.

Rogers, R. (2019). *Doing digital methods*. London, United Kingdom: SAGE Publications Limited.

Sampson, T. (2011). *Virality. Contagion Theory in the Age of Networks*. Minneapolis: University of Minnesota Press.

Shifman, L. (2013). *Memes in Digital Culture*. Cambridge, United States and London, England: The MIT Press.

Squirrell, T. (2017). Linguistic data analysis of 3 billion reddit comments shows the alt-right is getting stronger. Retrieved from https://qz.com/1056319/what-is-the-alt-right-a-linguistic-data-analysis-of-3-billion-reddit-comments-shows-a-disparate-group-that-is-quickly-uniting/

Tuters, M., & Hagen, S. (2019). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*. doi: 10.1177/1461444819888746

Tuters, M., Jokubauskaitė, E., & Bach, D. (2018). Post-Truth Protest: How 4chan Cooked Up the Pizzagate Bullshit. *M/C Journal*, *21*(3).

Van Soest, J. (2019). Language innovation tracker. Detecting language innovation in online discussion fora (Unpublished master's thesis), Vrije Universiteit Brussel, Belgium.

Willaert, T., Van Eecke, P., Beuls, K., Steels, L. (2020). Building social media observatories for monitoring online opinion dynamics. Social Media and Society. April 2020, doi:10.1177/2056305119898778

About the authors

Tom Willaert, Artificial Intelligence Lab, Vrije Universiteit Brussel. tom@ai.vub.ac.be
Paul Van Eecke, Artificial Intelligence Lab, Vrije Universiteit Brussel. paul@ai.vub.ac.be
Jeroen Van Soest, Artificial Intelligence Lab, Vrije Universiteit Brussel. jeroen.van.soest@vub.be
Katrien Beuls, Artificial Intelligence Lab, Vrije Universiteit Brussel. katrien@ai.vub.ac.be

Funding acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 732942.