Network toxicity analysis

An information-theoretic approach to studying the social dynamics of online toxicity

Kiddle, R. ; Törnberg, P.; Trilling, D.

DOI
10.1007/s42001-023-00239-2

Publication date
2024

Document Version
Final published version

Published in
Journal of Computational Social Science

Journal of Computational Social Science (2024) 7:305–330
https://doi.org/10.1007/s42001-023-00239-2

RESEARCH ARTICLE

Network toxicity analysis: an information-theoretic approach to studying the social dynamics of online toxicity

Rupert Kiddle[1] · **Petter Törnberg**[1] · **Damian Trilling**[1]

Received: 21 August 2023 / Accepted: 22 November 2023 / Published online: 1 February 2024

Abstract

The rise of social media has corresponded with an increase in the prevalence and severity of online toxicity. While much work has gone into understanding its nature, we still lack knowledge of its emergent structural dynamics. This work presents a novel method—network toxicity analysis—for the inductive analysis of the dynamics of discursive toxicity within social media. Using an information-theoretic approach, this method estimates *toxicity transfer* relationships between communicating agents, yielding an effective network describing how those entities influence one another, over time, in terms of their produced discursive toxicity. This method is applied to Telegram messaging data to demonstrate its capacity to induce meaningful, interpretable *toxicity networks* that provide valuable insight into the social dynamics of toxicity within social media.

Keywords Online toxicity · Telegram · Social media · Network analysis · Transfer entropy · Information theory

Introduction

The rise of social media has been associated with an increase in the prevalence of online toxicity, creating a clear risk for the health of civil discourse. Usually defined as "... the usage of rude, disrespectful, or unreasonable language that will likely provoke or make another user leave a discussion" [21, p. 3275], toxic interactions have become an undesirable fixture of contemporary online experience, with

✉ Rupert Kiddle
r.t.kiddle@uva.nl

Petter Törnberg
p.tornberg@uva.nl

Damian Trilling
d.c.trilling@uva.nl

[1] University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam 1080 WV, North Holland, The Netherlands

approximately 40% of US adults self-reporting as victims of online toxicity, and with severe encounters becoming increasingly prevalent [38].

Recent scholarship has investigated the dynamics of toxicity on social media by drawing inspiration from the study of the transmission of contagious disease. Combining epidemiological modelling with network analysis allows us to study the propagation of toxic behaviour across online platforms [21, 23]. These contagion-based approaches have shown that toxic behaviours are indeed transmissible, and that they spread across social media. However, while these approaches shed important insights into the dynamics of toxicity propagation, they are limited in their capacity to capture the nuances of individual-level causal effects. As the methods do not allow us to dissect the influence of individual agents over one another, we are limited in our ability to understand the flow of toxicity within social media, and the role of specific individuals in driving discursive toxicity. While we know much about the prevalence and harms of online toxicity, due to these methodological limitations, less is understood about how toxic behaviours emerge within social media and are shaped by the structure and dynamics of social interaction.

To address this research gap, this paper proposes a novel method— *network toxicity analysis* (*NTA*)—for the inductive study of the emergent dynamics of toxicity within social media. This method adopts an information-theoretic framework, based on a multivariate (conditional) transfer entropy (mTE) estimation procedure. This procedure yields an effective *toxicity network*, describing how entities within social media influence one another over time, in terms of their produced discursive toxicity. We term this their *toxicity transfer* relationship. This method does not depend on data on the underlying social network; relationships are inferred from entity-level time series data alone. By modelling the impact of individual entities on others' discursive toxicity, NTA provides a fine-grained understanding of how individual agents contribute to the spread or mitigation of toxicity on social media. Moreover, it provides a means to locate the specific drivers of toxicity within a network, enabling targeted interventions and other strategies to address the growing problem of toxic behaviour within online social media platforms.

To demonstrate NTA on real-world social media data, we apply it to a large-scale Telegram dataset. This dataset is a 6-month snapshot of the Dutch-language 'Telegramsphere' (i.e. public chats and channels). We examine the capacity of the method to yield an interpretable *toxicity network* that describes the emergent structure of discursive toxicity within this relevant public discussion space. With the aid of established topological measures, we identify key structural drivers of toxicity at the network and neighbourhood levels, and complex and multi-scale interactions between network elements underlying the transmission of toxicity.

Background: modelling online toxicity

Scholarship has demonstrated that toxic online behaviours have far-reaching consequences. Firstly, these behaviours have been found to significantly impact the well-being of individuals targeted by them, leading to increased stress, anxiety, and even depression [8]. Secondly, toxic behaviours reduce the perceived social utility

of online exchanges by fostering an atmosphere of negativity, hostility, and distrust [3, 25]. When toxic interactions dominate online spaces, users may feel discouraged from participating in discussions, sharing their perspectives, or seeking information, thereby undermining the potential benefits of online platforms as spaces for civic collaboration and knowledge sharing. Finally, toxic discourse has been shown to selectively poison sensitive social and political discourses, hindering meaningful conversations and contributing to the polarization of opinions within society [28].

The detrimental impacts of toxic behaviour exceed those of individual interactions, as toxicity appears to be socially transmissible, with toxic performance being driven by exposure to toxic [15] or 'triggering' language [1]. Exposure to toxic online interactions has thus been shown to influence individuals to adopt similar behaviours. This phenomenon, known as social contagion, occurs when people observe and imitate the actions and language of others within their social network [7]. Social media platforms, with their wide reach and instant connectivity, provide fertile ground for the rapid transmission of toxic behaviour: when users witness or participate in toxic exchanges, they may internalize and replicate those behaviours in their future interactions, thus perpetuating a cycle of toxicity. As a consequence of this, toxicity has been found to be remarkably prevalent on social media compared with other online domains [38]. The contagious nature of toxicity on social media is thus a pre-eminent concern, as it not only threatens the quality of online discourse but also creates an immense regulatory burden for platforms that must combat hateful and harmful language through large-scale moderation and policing.

To address the socially transmissible nature of toxicity within social media, researchers have recently turned to contagion models that integrate agentic theories of disease-spreading from behavioural epidemiology with social network analysis—a widely used tool for examining emergent structural effects in online spaces [5]. These contagion approaches categorize social media users into pre-defined agent states, which—in a minimal model—represent whether they have exhibited toxicity or not (i.e. whether they have become 'infected'). They analyze temporal changes in these states as users are exposed to content with varying levels of toxicity. Essentially, these models study the contagiousness of toxicity across a population of online agents and attempt to provide a best fit for such contagion by varying the properties—as informed by underlying theories of behaviour—of those pre-defined agent states.

These contagion approaches have revealed that toxicity should indeed be understood as a socially transmissible phenomenon [21, 23]. However, while these methods have been valuable in testing and refining theories of agentic behaviour to describe toxicity propagation on social media, they do not offer the type of detailed empirical understanding that would allow developing a more comprehensive understanding of the emergent structures of toxic influence within social media. To examine the details of the dynamics of social interaction, it would be necessary to also capture the causal interaction of specific agents responsible for driving the discursive toxicity of others. These limitations of the current contagion-based approaches stem from significant assumptions made regarding (i) the scope of agent behaviour and (ii) the structure of potential influence, resulting in a limited ontology of communication dynamics.

First, contagion approaches to modeling the spread of toxicity impose a predefined model of behaviour, i.e, a limited set of agent states, which restricts their capacity to reflect complex patterns of behavioural propagation. By constraining behaviour to predetermined states (e.g. susceptible, infected, or recovered), contagion models are unlikely to capture the nuanced interplay between personal characteristics, social contexts, and other situational factors that influence the adoption of toxic behaviours. As such, they may fail to reflect the full multifaceted nature of online toxicity dynamics.

Second, the imposition of a predefined network structure means that contagion approaches provide a restricted ontology of influence, and this has downstream consequences for their capacities to provide an accurate picture of toxicity propagation within social media. By assuming predetermined pathways of toxicity spread based on observable connections such as friendships or followerships, contagion models overlook the potential for indirect or unobserved pathways of influence, such as cross-domain interactions or offline relationships. This imposition of structure furthermore assumes equal influence amongst connections, disregarding the varying degrees of influence individuals may have on one another in reality [2].

Dependency on a predefined network structure also results in practical limitations. Contagion models often struggle to be applied in situations where there is a scarcity of relational data. This is particularly evident in the case of 'dark platforms' [26, 32, 40], where researchers face restrictions in accessing data about users' connections, such as friendships or follower networks. The challenges of studying 'dark networks'—where relational data are limited or unreliable—are well-known within domains such as criminology [6, 10, e.g.,], but are relatively novel to the study of social media. While researchers have often taken it for granted that such data would be available via APIs, the demise of freely accessible APIs as well as the expansion of data protection and privacy legislation makes it increasingly problematic to rely on such assumptions. These limitations highlight the need for more flexible modeling approaches that can account for diverse pathways of influence and adapt to data constraints in order to enhance our understanding of toxicity propagation in social media.

In summary, contagion approaches have demonstrated the centrality of transmission in the dynamics of online toxicity, and provided valuable insights into the spread of toxicity across social media. Their main strength lies in their ability to compare theories of agentic behaviour and understand system-level dynamics when accurate data describing the underlying relational structure is available. However, these methods are limited by their restricted ontology and the increasing difficulty of obtaining necessary data. This suggests the need for a complementary method that can infer toxicity influence relationships inductively based solely on observed behaviour. In the next section, we propose such a method.

Framework: network toxicity analysis

Our approach to studying the dynamics of online toxicity combines an information-theoretic technique (multivariate transfer entropy, or mTE) with established network analysis tools. We use the former to model the *transfer* of toxicity from one

communicating agent to another within social media data. A communicating agent is understood as an entity that exhibits some measurable toxic performance over time. A *transfer* of toxicity between two communicating agents is understood as a significant reduction in the uncertainty about a given target's *future* toxic performance that is afforded by having knowledge of a given source's *past* toxic performance, whilst conditioning on the past of the target itself as well as those of all other known communicating agents in the system (i.e. in the social media data). This multivariate approach yields a *toxicity network*, which is a spatial description of all *toxicity transfer* relationships observed within the system. In other words, it describes who influences who, in terms of their produced discursive toxicity. Since this is a network, we will use the terminology of network analysis to discuss the method. As such, a 'node' represents a communicating agent, denoted as n_i, while an 'edge' symbolizes the toxicity transfer relationship between two nodes, denoted as e_{ij}, where i and j correspond to the indices of the connected nodes.

NTA extends existing information-theoretic approaches to modelling influence on social media. Building upon foundational theoretical work by Schrieber [29], Ver Steeg and Galstyan [33] introduce a method for estimating 'information transfer' within social media data. They propose a bivariate measure of time-dependent information transfer between communicating agents within a network of unknown structure. Their approach adopts an information-theoretic technique called *transfer entropy* [29], which takes two stochastic processes and estimates the reduction in uncertainty about one process implied by having knowledge of the other process. They demonstrate the capacity of their approach to capture influence on a fine-grained level, to reveal meaningful and otherwise hidden network structures (a finding also validated by Bauer et al. [2]) and to provide an easy to digest, predictive interpretation. Their follow-up work further developed this approach to numerically incorporate the textual content of communications, yielding a measure of 'content transfer' that was used to estimate the degree of influence that one Twitter user had over the topical content of another [34]. This allowed them to identify important pair-wise influence relationships within Twitter messaging data based on the textual performances of those users alone.

In NTA, we draw on the key innovation of this work—the pairing of a numerical representation of text with a pairwise transfer entropy estimation technique—and translate it into a multivariate framework that is able to yield an estimation of a complete 'influence network'. This translation is necessary to address a limitation of their approach. Whilst they detected important pair-wise relationships, some of these may have been spurious or redundant in cases where multiple source nodes provided identical information about the target (false positives) and others might have been underestimated due to missed synergistic interactions between different source nodes, resulting in greater total transfer (i.e. influence) to a target node (false negatives). A multivariate estimation technique circumvents these issues, proving a clearer picture of the structure of influence [19].

To achieve this, we adopt the multivariate (also called conditional) transfer entropy (mTE) estimation technique encoded within the open-source Python package: Information Dynamics Toolkit xl [39]. This approach allows for the efficient estimation of mTE. Ordinarily, doing so via a 'brute force' method—with even a

relatively small number of nodes—quickly becomes computationally intractable. The IDTxl algorithm circumvents this by instead approximating mTE using a 'greedy' approach that iteratively assembles a non-uniform embedding—Z—constituted by the selected pasts of all relevant source variables on the basis of their contributing maximum, significant and non-redundant (in the context of the final embedding set) information about the next state of the target variable [9, 19, 39]. We apply this estimation procedure to quantify the edgewise informational contribution—specifically, the reduction of uncertainty—that the knowledge of the past toxicity exhibited by a given source node ($X^{-}_{n_i}$), provides about the future toxicity exhibited by a target node, ($Y^{+}_{n_j}$). This is achieved by conditioning not only on the past of the target node itself, ($Y^{-}_{n_j}$), but also on the pasts of all other pertinent source nodes within the network. The latter is encapsulated by the non-uniform embedding, represented by Z, which includes the relevant historical data from the entire set of nodes influencing Y_{n_j}. Whenever the past of a source provides statistically significant information about the future of a target, this indicates that the source has influenced the target. We term this a *toxicity transfer*.

This estimation process yields an approximation of the underlying *effective* [4, p. 143] toxicity network; that is, an inductive description of the *time-dependent* and *conditional* toxicity transfer relationships between network elements. In other words, it allows the researcher to reasonably estimate the degree of toxic influence that a collection of communicating agents have over each another, based only on observations of their time-dependent behaviours, and in circumstances where we know nothing of their pre-existing relationships. Importantly, the toxicity transfer relationships revealed by this method are not strictly causal. Rather, they express a measure of information transfer between nodes that ultimately rests on (conditional) correlation, and thus the induced *toxicity network* will capture exogenous influences that exceed the endogenous causal capacity of the system (i.e. the available social media data) being observed [20]. However, given that in the study of social media—and indeed for the study of toxicity therein—the possibilities for experimental perturbation or intervention are often scarce, the mTE method provides an exceptionally valuable means of *approximating* causal networks based on easily-accessed observational data only. To the best of our knowledge, this work constitutes the first application of an mTE framework to model content dynamics (influence) within textual data.

Components of the NTA approach

There are four components to NTA: (1) the input data, (2) toxicity measurement, (3) process re-sampling and (4) toxicity network estimation. What follows is a high-level overview of the methodology. A toy example is provided in Fig. 1. An applied methodology, tailored to a specific platform (Telegram), will follow in the next section. A repository containing code (in Python) for the general application of NTA to social media data is also made available [14].

For input data (1), NTA requires a time-ordered, entity-attributed, text-based content stream. For example, this may be a stream of messages or posts,

for which the sender or forum is uniquely identifiable, and with sufficient time and date information available such that the ordering of communications can be established unambiguously.

To classify toxicity (2), messages are converted to numerical representations of their respective toxicity level. For this task, we use Google and Jigsaw's Perspective API [11], which returns a probabilistic score, but any measurement approach yielding a sufficiently fine-grained continuous result is feasible.

In the following step (3), each entity vector of toxicity scores (i.e. the complete toxic performance of a single communicating agent) is re-sampled so that all have an equal sampling rate. This is a requirement of IDTxl's estimation algorithm as it was implemented at the time of publication. The re-sampling strategy should be tailored to an understanding of the expected causal dynamics of the social media platform under study (see the next section for an applied example). The result of

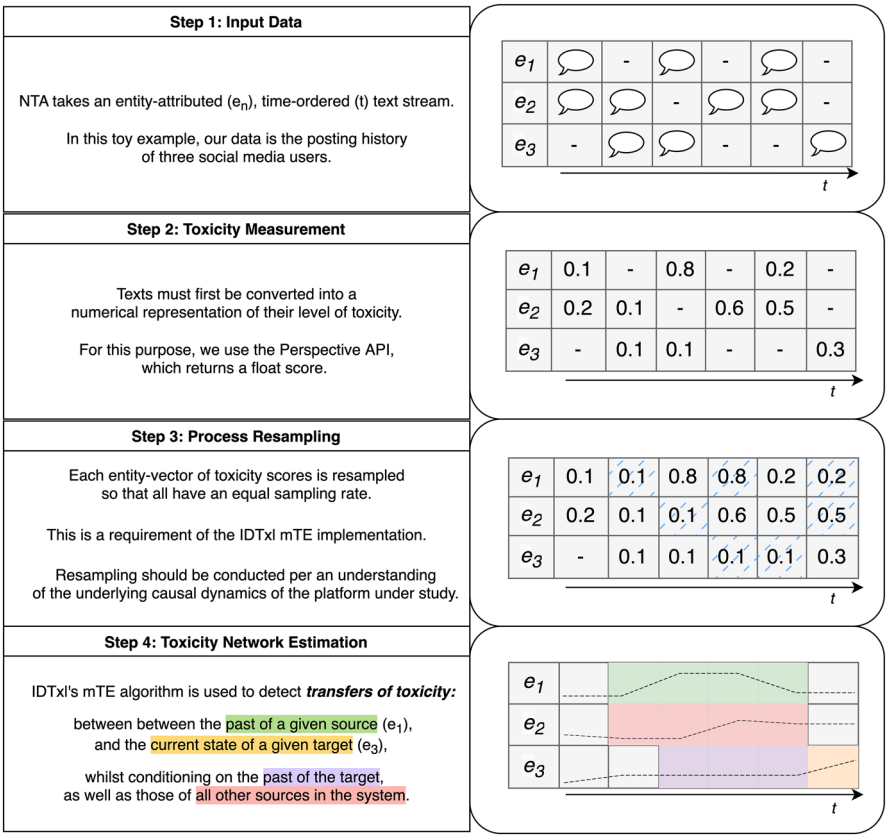

Fig. 1 A toy example of the NTA method applied to time-series data for three fictitious social media users. For detail on the toxicity measurement, see [11]. For further information on the mTE estimation procedure, see [39]

this step is an array of equally sampled toxicity vectors, with each representing the level of toxic performance exhibited by a single agent over time.

To estimate network toxicity (4), the resultant array of re-sampled toxicity vectors are provided as input to IDTxl's mTE algorithm to estimate the frequency with which statistically significant toxicity transfers took place between observed agents. To do this, processes must first be chunked into time-limited slices. The length of these slices should be as short as possible (since the number of slices determines the maximum number of times that a transfer event can be detected; more slices allows for clearer differentiation between estimated relationships) while remaining long enough to provide sufficient information for detection of linear or non-linear effects, depending on the estimator settings used. Key configuration options include: (i) the choice of linear (less sensitive, but faster) or non-linear (more sensitive, but slower) estimator, (ii) specification of the minimum and maximum lags (how far in the past to measure transfer effects within each slice), and (iii) critical alpha levels for all statistical tests. Once estimation is complete, the slices are aggregated to produce a toxicity network, where each node (n_i) represents a communicating agent and each edge (e_{ij}) represents a *toxicity transfer* relationship between n_i and n_j, weighted by the (relative) frequency with which statistically significant transfers of toxicity took place (w_{ij}).

Attributes of the NTA approach

The NTA approach to studying toxicity propagation differs fundamentally from a contagion-based approach in that it is 'flow-based' rather than agent-based. Whereas the latter essentially asks which combination of agent properties (states) best describes the onward transmission of toxicity given a predefined (known) network structure, this flow-based approach induces the temporal relationships between toxicity 'streams' as they emerge from *potentially* communicating entities within a network of an undefined structure. This abstraction away from agents and towards information-flows represents a shift towards studying online environments as complex ecosystems, comprising mixed social and technical elements that interact at different scales to produce emergent outcomes.

The 'model-free' [33, 34] nature of this approach allows for its application in situations where underlying network structures are unknown, making it particularly valuable when examining platforms such as Telegram or WhatsApp, where access to relational data is limited. It surpasses the constraints of pre-defined network models by identifying potential influence relationships that might go unnoticed by contagion methods, including unobserved or offline interactions that exist beyond direct connections. Furthermore, its abstention from imposing a model of behaviour (i.e. agent states) results in an inductive instrument that is sufficiently receptive to 'surprise' such that it can detect the outcomes of complex interactions at different scales of social operation, as well as the interaction between social and technical elements (affording a non-anthropocentric epistemology). This allows the researcher to move between these scales as demanded by the underlying research interest.

Even in the absence of structural inputs, this approach should still reflect the outcomes of social reinforcement mechanics. Recent evidence [24] suggests that an information-theoretic approach to understanding the propagation of information over social networks is sensitive to complex contagion dynamics (such as the weakness of long ties and slowed transmission in dense areas) despite lacking an explicit mechanism for social reinforcement. As such, NTA should appreciate these dynamics. This is important since the traversal of controversial social and political information across social media displays complex contagion-like dynamics [22]. As toxic messages are often controversial by nature, these dynamics are arguably essential to understanding the propagation of toxicity on social media.

Demonstration: Telegram messaging data

We use the instant-messaging platform, *Telegram*, as a case for two key reasons. First, being an exemplar of a 'dark platform' [26, 32, 40], it is a context for which we lack reliable structural data with which we could otherwise model the propagation of toxicity using other means (such as a contagion model). On Telegram, the vast majority of discussion fora that constitute the 'Telegramsphere' (all non-private Telegram chats and channels) severely restrict the availability of user data. Chats (many-to-many fora) may or may not make their participant lists visible, and channels (few-to-many fora) do not publish their follower lists. In addition, users may change their handles at any time. These factors mean that the underlying structure of association remains largely unknown to researchers of this platform. Applying NTA to Telegram data demonstrates the utility of the method in inferring structural relationships in the partial to total absence of such data.

Secondly, Telegram is an exceptionally low moderated and permissively governed platform that has seen much of its recent growth stem from the arrival of those who have been 'deplatformed' from mainstream counterparts [27]. It is—in many contexts—the platform of choice for the far-right, conspiracy theorists, hate groups and, more generally, those whose message would be deemed unacceptable and subject to censure on mainstream platforms [12, 17]. It is therefore a relatively volatile public discussion space, arguably more at risk of generating toxicity amongst its users. It is therefore valuable to study the dynamics of toxicity within, and to identify which communities stand out as toxic influencers, driving the discursive toxicity of other communities and potentially leading to their further radicalization.

Our data consists of a 6-month slice taken from a larger 4-year snapshot of the Dutch 'Telegramsphere' (Dutch-language public discussion chats and channels) collected by Simon et al. [32]. This slice spans the period of January to June 2021, consisting of (i) message text, (ii) forum (the name of chat/channel where the message was posted) and (iii) date and time of posting. This selection is based on the relative and sustained increase in messaging activity beginning at the start of this time period (see Fig. 2).

Whilst representing only a small portion of the overall timeline, it captures over 75% of the messaging data (around 1.6 million messages) and includes 93% of the channels ($n = 162$) in the full dataset. It is also a substantively interesting period of

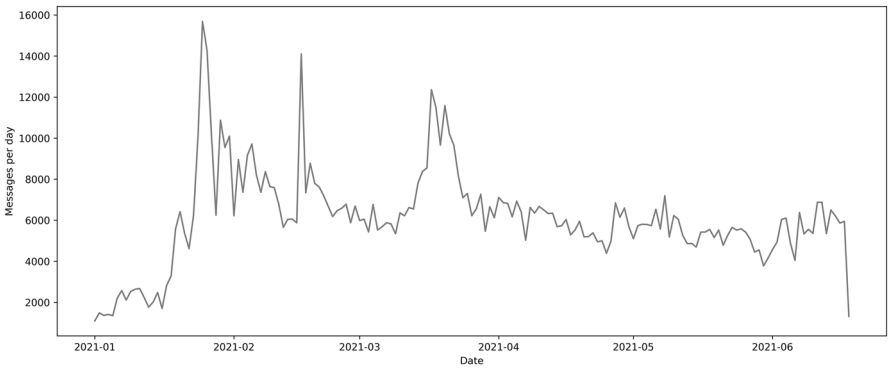

Fig. 2 Number of messages sent per day across all ($n = 162$) chats/channels

time in that it corresponds with the period of extended 'hard lockdown' in the Netherlands that was enacted in response to the COVID-19 crisis. This began in January 2021 with the announcement of the 'avondklok' (a nighttime curfew that lasted for almost three months) and ending just prior to the lifting of many social restrictions in June with the 'openingsplan'. This was a period of substantial civil unrest with elements of the far right and fringe taking simultaneously to the streets, as well as to Telegram itself, to organize rallies and riots to sometimes violently protest lockdown measures [32]. It also featured the Dutch general elections, which took place mid-March at the height of social restrictions. It therefore stands as a potential hotbed of toxic influence.

We treat chats (many-to-many fora) and channels (few-to-many fora) as communicating entities rather than individual users, for reasons both substantive and practical. Substantively speaking, in the context of Telegram, it makes sense to describe toxicity transfer relationships at the level of messaging communities, since these are self-sorting communities. Users self-select into chats and channels on the basis of existing relationships (connections) and on the basis of what sorts of content they want to see (preferences), in a context relatively free from algorithmic interference. As such, a focus on messaging communities as communicating entities allows us to address whether and to what extent far-right and other fringe groups produce toxicity and to what extent this toxicity influences that of other groups. Additionally, it allows us to make use of a relative freedom offered by this approach: since we assume nothing of the underlying dynamics, we are free to move between different scales of analysis as required. In practical terms, taking chats/channels as communicating entities significantly reduces computational demands, minimizes issues of data sparsity, simplifies interpretation, and allows for substantive discussion without risking identification of individual users.

Applying NTA

We first converted individual messages into numerical representations of their relative level of toxicity, using the Perspective API. Each message was sent to the API

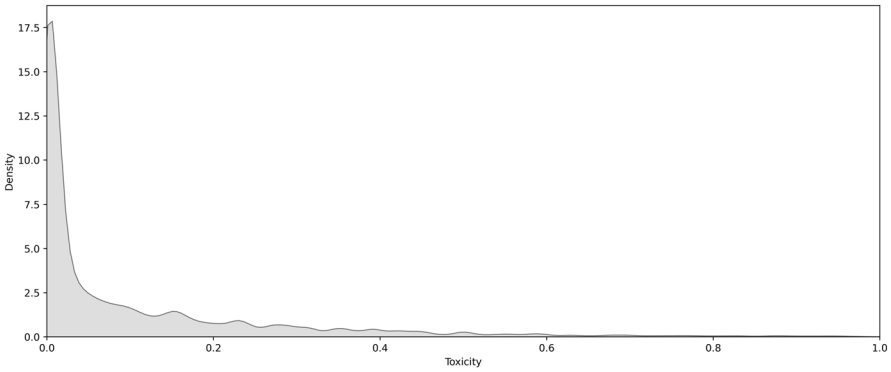

Fig. 3 Kernel density estimate plot describing the overall distribution of toxicity scores

as UTF-8 text, and a score was returned for each message (a float between 0 and 1) that represented the probability that a reader would perceive the message as being toxic as per the given definition: "[a] rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" [21, p. 3275]. Classification was successful for ≈ 82% of messages; or around 1.3 million. The distribution of toxicity scores (Fig. 3) was heavily weighted towards zero (meaning that most messages were probably not toxic) but with a long tail (meaning that a relative minority of messages varied substantially in their probability of being toxic). On average, a message had a 10% probability of being considered toxic by a reader ($SD = 16\%$).

Following conversion of individual messages into toxicity scores, each chat/channel in the dataset was represented as a point-process-like vector of toxicity scores. To achieve an equal sampling rate, we developed a re-sampling strategy based on the expected causal dynamics of in-platform message visibility. We re-sampled for 5-min intervals, by using two conditionally triggered rules. For the first rule: if a 5-min window contained $\geqq 5$ messages, we took the simple average of the toxicity scores that fell within that window. The logic here was that for high-frequency periods, messages fade quickly from visibility and thus from causal relevance; thus taking a simple average captures maximal information while remaining respectful of the temporality of message visibility. For the second rule: if a 5-min window contained < 5 messages, we took the simple average of up to the last 5 messages, so long as they fell within the previous 24 h. The logic here was that for low-frequency periods, messages remain causally relevant for up to 24 h as they remain visible at the top of a chat/channel to those users that only check their app once per day. If no scores were found within the past 24 h, NaN was assigned.

With each chat/channel in the form of an equally sampled toxicity vector, we then passed this data to IDTxl's algorithm to estimate the effective network. To do this, $n = 495$ sequential slices of 8 h (N = 00:00 to 08:00, AM = 08:00 to 16:00, PM = 16:00 to 00:00) containing 96 samples each were created and passed to the estimator sequentially; chats/channels were only included in each slice if they had no missing data points (after re-sampling) for that particular period.

For the mTE calculations, a Gaussian conditional mutual information estimator was used. This has the limitation of only being able to observe linear relationships, so it is likely to have detected a far sparser set of transfer relationships than would be seen with a non-linear estimator (e.g. with a Kraskov estimator; [16]). Additionally, parameters for minimum and maximum sample lag were set to 1 and 6 respectively, meaning that the estimator could only detect transfers of toxicity between chats/channels occurring within the previous half-hour ($T - 30$) within each slice. These two limitations (linear estimator, half-hour search window) will be returned to in our discussion since scaling of this workflow could enable a more comprehensive analysis.

The estimation procedure yielded network data for $n = 495$ sequential slices. This consisted of edge lists detailing pairs of chats/channels where statistically significant transfers of toxicity took place. We aggregated these by first summing the counts of significant transfers across all slices edge-wise, and then dividing these counts by the maximum number of times that any single edge was included in an estimation ($n = 449$). The resultant toxicity transfer metric was thus comparable across edges that may have been included in different numbers of estimations due to differing degrees of missing data after re-sampling (edges with missing data were 'punished' proportionally). This final metric, therefore, described the relative degree of toxicity transfer between any n_i, n_j over the 6-month observation period.

As a final step—and in order to establish a clearer picture of the underlying effective network structure—we extracted a simplified 'backbone' of the most important edges in the network, using a statistical method that preserves complex multi-scale interactions [30]. After filtering ($\alpha < 0.05$), only 3% of edges were retained in the final network, constituting the most important toxicity transfer relationships.

For all other statistical tests used in the estimation process, the critical alpha level was set to $\alpha < 0.01$, with the number of permutations set automatically in order to satisfy this level. Due to the large number of nodes involved in each estimation, it was not feasible to perform false discovery rate corrections in most cases. This likely led to some degree of overestimation (false-positives) in the raw effective network data returned by the estimator. However, since we subsequently apply an aggressive statistical backbone filter (97% of edges removed) to the result, the effects of this on the final network are presumed to be marginal.

Results of NTA

Figure 4 shows the *toxicity network* produced from our analysis of the 6 months' worth of Telegram messaging data. Chats (shown as circles) and channels (shown as squares) make up the individual nodes (n_i) in the network. The size of a node indicates its relative importance in influencing toxicity throughout the whole network, as measured by the PageRank method [13]. Colour indicates its relative level of toxicity (red is more toxic), measured as the proportion of messages that scored a 50% chance or higher of being toxic. For a full description of the results for every chat and channel, please see Appendix B.

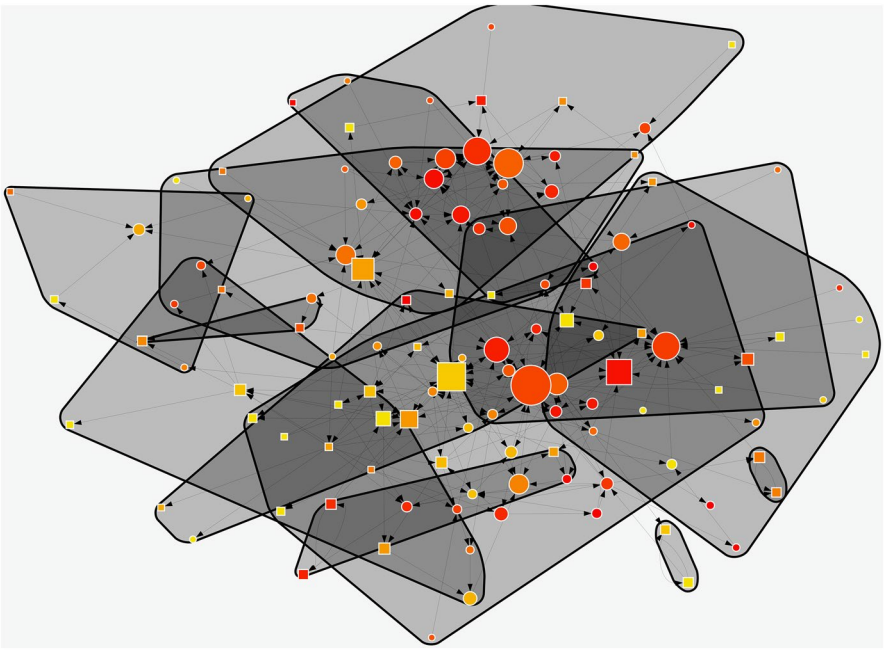

Fig. 4 Structure of toxicity transfer relationships within the Dutch-language Telegramsphere, Jan–Jun 2021. Node shape denotes type: chat (circles) or channels (squares). Node size indicates network influence (larger is more influential) measured by PageRank. Node colour indicates toxicity (red is most toxic) measured by the proportion of messages that scored a 50% chance or higher of being toxic. n = 118 chats and channels (44 isolates not included). See Appendix B for further detail

Edges (e_{ij}) represent the presence of a persistent and significant toxicity transfer relationship between two chats/channels in the direction indicated (bidirectionality is frequent). The presence of an edge indicates that the past 30-min' worth of toxicity measurements from the source frequently held significant predictive information about the future toxicity of the target, while taking into account (controlling for) the future effects of the pasts of the target itself as well as all other nodes in the network. In other words: an edge indicates that toxicity was regularly transferred from source to target, constituting (conditional) toxic influence. Each edge weight (w_{ij}) represents the relative strength of this relationship (essentially: how robust it was over time). These range from 0.010 to 0.093, with a mean ($\bar{x}$) of 0.040 and a standard deviation (SD) of 0.018; they can be understood as percentages, reflecting how often a given directed edge was involved in significant toxicity transfer as a proportion of the maximum number of times that any edge was considered in the estimation process.

Not shown in these networks are those chats/channels ($n = 44$) that had no significant toxicity transfer relationships after the pruning process was completed (isolates). This does not mean that they had no influence over the toxicity of others, only that the strength of that influence was not significantly higher

(with $p = 0.05$) than what would be expected given a null model for the local assignment of weights to edges [30]. These will not be treated analytically since the pruning process essentially renders them as one homogeneous group, masking underlying heterogeneity. Instead, our analysis focuses on what remains after their removal: the most important (statistically speaking) toxicity transfer relationships and the chats/channels that drove these relationships over the 6-month observation period.

We observed a toxicity network of 118 chats/channels sharing 302 edges. This was a sparse network (*density* = 0.02) with two influential toxicity *exchanges* (located mid-center and upper-center), as well as various smaller formations, indicating the presence of different scales of interconnected toxicity-producing communities. The essential interpretation of this network is that nodes that are larger and more central were more important in driving toxic discourse within it. They in other words act as what we would commonly refer to as *toxic influencers*, and drive the overall toxicity of the larger network.

Community detection using the info-map method reveals several fairly well-defined (*modularity* = 0.42) communities that overlap with one another substantially. Degree assortativity is low-moderate and negative ($r = -0.26$), indicating that nodes with a higher degree tended to connect preferentially to those with a lower degree. We find similar, albeit weaker, association results for neighbourhood-level influence, as measured by the sum of outward toxicity transfer scores for each chat/channel ($strength_r = -0.14$). Finally, we observe essentially random associations on the basis of network-wide influence ($PageRank_r = 0.002$). Together, these observations indicate a well-connected and mildly disassortative underlying structural network, consistent with conclusions drawn by Simon et al. [32] in their own methodologically distinct analysis of this data. These qualities are conducive to the diffusion of toxicity between chats/channels and consistent with the notion of toxicity as a pervasive and ubiquitous phenomenon on social media [38].

The average level of toxicity per chat/channel appeared to be distributed randomly. In other words, chats/channels did not appear to cluster together on the basis of their toxicity. An assortativity of −0.10 revealed that this was generally but not always the case: higher toxicity chats/channels did display a slight tendency to connect preferentially with lower toxicity counterparts. Additionally we saw no clear visual association between toxicity (colour) and influence (size). This implied that the toxicity of a chat/channel and its degree of influence over the toxicity of others were unrelated to one another. Furthermore, we found no significant relationship between the level of toxicity (measured either as a simple average, or as the proportion of scores greater than 50%) and either the PageRank (whole-network influence) or strength (neighbourhood influence) scores. In sum, these observations allow us to conclude that the level of toxicity of a chat/channel appears to be unrelated to its ability to influence the toxicity of other chats/channels.

Figure 5 introduces the hand-coded topical classifications for each chat/channel [32, pp. 49–50]. We observed a discussion space dominated by non-mainstream groups: the far-right, alternative news sources, activism groups, COVID-skeptic and conspiracy-themed chats and channels. However, the distribution of these communities was not random ($r_{topic} = 0.090$), and we observed the emergence of two distinct

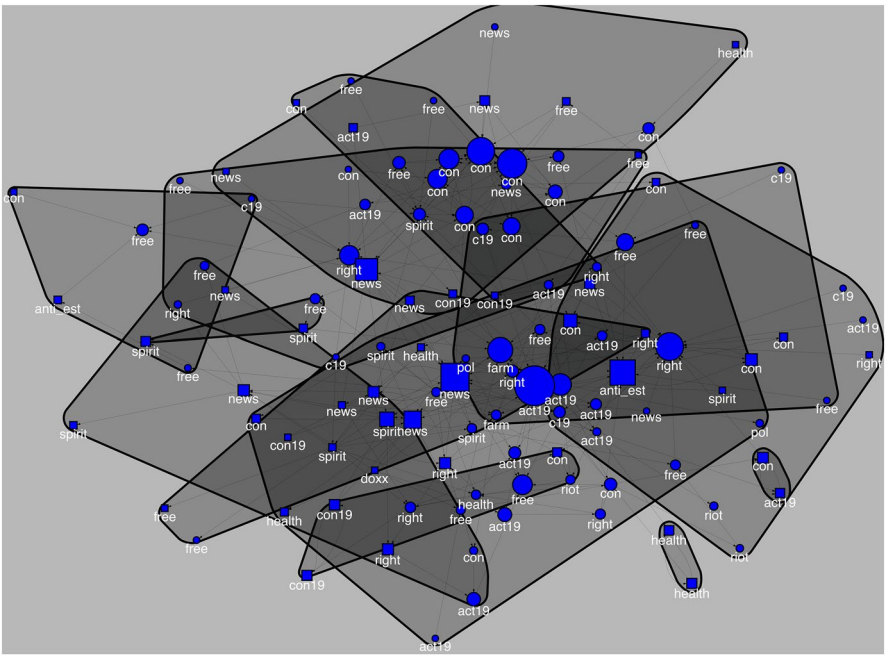

Fig. 5 Structure of toxicity transfer relationships within the Dutch-language Telegramsphere, January to June 2021. Hand-coded topical categories for all chats/channels from Simon et al. [32], see Appendix A for codebook. $n = 118$ chats and channels (44 isolates not included). See Appendix B for further detail

'toxicity exchanges'. Observing the mid-center exchange, we found a region of mixed chats and channels whose foci concerned the political right, alternative news, anti-COVID-measures activism, freedom activism and health/spirituality. This was a (relatively) diverse cluster, both in terms of forum type (chat/channel) and topic, that constituted a central point of toxicity generation and exchange within this network. It included the most notable public authority within the dataset, the official channel of the Dutch right-wing political party 'Forum voor Democratie (FvD)'. The interpretation afforded by NTA is that this relatively diverse exchange can be considered central to, and highly influential in, driving flows of discursive toxicity within the Dutch Telegramsphere over this 6-month observational period.

Moving to the upper-center exchange, we observed a cluster of conspiracy-themed chats. These included the notable conspiracy group 'DeBataafseRepubliek' that was temporarily taken offline by court order of the Public Prosecutor's Office in the Netherlands for its role in disseminating dangerous conspiratorial content [18]. This exchange is distinct from the last in that it presents as highly homogeneous in terms of both topic (conspiracy) and forum type (chat). This topology indicates that these conspiratorial chats not only fueled each other's toxic discourse but also, due to their significant network influence, extended their impact on discursive toxicity to the wider network of non-conspiratorial chats and channels. This observation could provide empirical evidence of an emergent 'feedback loop' [36]. This homogeneous network of discussions, characterized by their conspiratorial themes and

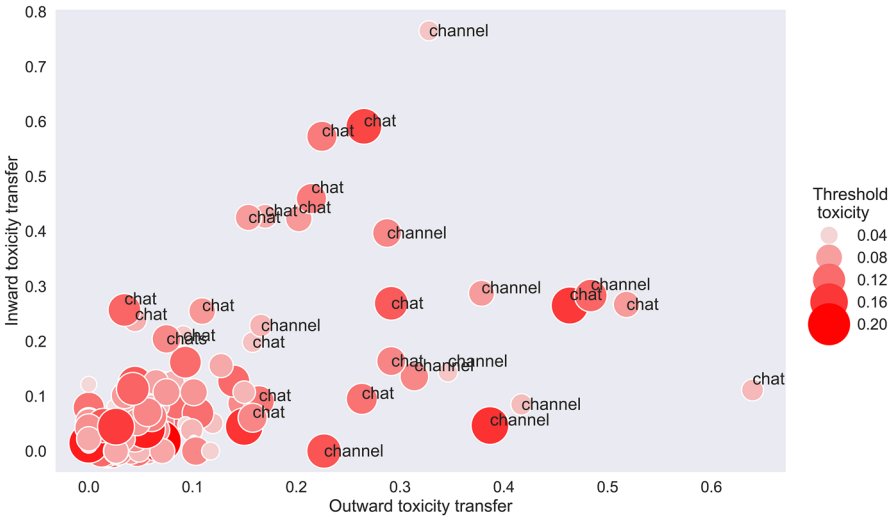

Fig. 6 Local influencers and followers within the Dutch-language Telegramsphere. Axes are the sum of inward/outward edge weights from each node. Threshold toxicity describes the proportion of messages per channel with a higher than 50% probability of being toxic

peer-to-peer authority structures, appears to synergize, escalating the level of discursive toxicity and potentially creating a bandwagon-like effect [37] that amplifies their toxic influence within the broader discursive network.

So far, we have addressed chats/channels on the basis of their topographical locations within the network and on the basis of their degrees of influence over the discursive toxicity of the entire discussion network (via the PageRank method). However, we may have reason to want to address structural roles in toxicity transfer at a neighbourhood level. Figure 6 illustrates one basis for classifying chats/channels in such a way. We observe that the majority of chats/channels fall into the bottom-left quadrant (meaning that they are relatively less implicated in toxicity transfer relationships), but that a sizable minority of outliers stand out. These outliers constitute fora that are notable for their capacity to *either* influence or be influenced, in terms of discursive toxicity, at the level of the neighbourhood (i.e. all those chats/channels with which they share direct transfer relationships). Those that fall into the bottom right are what we might call local *influencers*: they exert a relatively large outward influence on other chats/channels in their neighbourhood. On the other hand, those that fall into the top left are what we might call *local followers*: their level of discursive toxicity appears to be highly driven by others within their neighbourhood.

We observe some differentiation of these roles in terms of forum type: channels tend to be influencers and chats tend to be followers. These tendencies are arguably consistent with their encoded authority dynamics: channels, as few-to-many fora, encode hierarchical authority within their communicative dynamics, whereas chats—as many-to-many fora—encode distributed authority. This suggests that the divergent socio-technical dynamics encoded by these two types of fora might have consequences for the dynamics of toxicity propagation on Telegram. Specifically,

channels might be more important than chats in driving toxicity at the local level (the neighbourhood). Importantly, this differentiation disappears when we consider network-wide influence via PageRank. This suggests that the structural drivers of discursive toxicity on Telegram operate at different interacting scales, and that the comprehensive identification of 'toxic influencers' demands an investigation into these varying scales of influence.

Discussion

In the previous section we demonstrated the capacity of NTA to estimate an effective *toxicity network*, describing the conditional *toxicity transfer* relationships between communicating entities (chats and channels) within a real-world (Telegram) dataset. When applied to these data, NTA was able to reveal meaningful and interpretable relationships between these entities in the complete absence of structural data describing the underlying relationships. Combining an information-theoretic inductive technique with the established tools of social network analysis, it allowed for the identification of the central drivers of discursive toxicity within the network. These included emergent communities (*toxicity exchanges*) as well as individual nodes (*toxic influencers*) that appeared to be responsible for driving neighbourhood-level and/or system-level network toxicity. The identification of agents within social media not only on the basis of their toxicity but on that of their ability to influence the toxicity of others is a particularly valuable capacity, inasmuch as it can transform interventional efforts to curb the spread of toxicity on social media by affording a dynamic and effect-oriented understanding of toxic online behaviour.

Our findings suggest that the level of toxicity exhibited by a discussion community does not necessarily predict its influence over the discursive toxicity of other (connected) communities. However, since this analysis was performed at the level of chats/channels on Telegram, further study is required in order to investigate the dynamics of individual (i.e. person-level) toxic influence, and in other online contexts (i.e. other platforms). In addition, our results suggest that local interactions within emergent communities can have system-level consequences for the structure of toxic discursive flows. Where the homogeneity of a community potentiated heightened local exchange, this appeared to produce an enhanced capacity to drive discursive toxicity within the greater network. In practical terms, this suggests that homogeneous communities (so called 'echo-chambers' [35]) might pose a greater toxic risk to online spaces if they act as generative toxic centers that then spillover into the broader discursive network. Since NTA is equipped to appreciate multi-scale reinforcement dynamics, it can be used to locate communities of concern that are synergistically implicated in network-level toxicity flows.

We also found evidence that suggests that the socio-technical dynamics encoded in platform design may have consequences for the spread of toxicity within social media. In the case of Telegram, we observed that channels (few-to-many fora that encode a hierarchical authority dynamic) are more likely to act as *local influencers* than chats (many-to-many fora that encode a distributed authority dynamic), which are more likely to act as *local followers*. The divergent social dynamics encoded

within Telegram's forum types are therefore arguably relevant for the flows of discursive toxicity within the platform. Similar effects may be observed on other platforms where design features encode authority-minded dynamics (e.g., verification badges). NTA allows for the evaluation of their effects over the flow of toxic behaviours within the platform.

Limitations

This paper presents a first application of a novel technique for the study of toxicity dynamics on social media, and as such, carries several limitations. These can be divided into technical and procedural limitations. By technical limitations, we refer to those that stem from the limited computational resources with which the demonstration with Telegram data was conducted. These included the (i) use of a linear estimator and the (ii) restriction of the causal search window to a half-hour. The linear estimator meant that only linear relationships were detected. This would have resulted in a far sparser network than would be possible with a non-linear estimator such as the Kraskov estimator ([16], also implemented within the IDTxL suite). The restricted search window meant that only short time-scale transfers (occurring within 30-min) could be detected. As such, the procedure likely underestimated the actual toxicity network to a degree that is difficult to estimate, since we lack knowledge about how quickly toxicity spreads over social media. Future research should mitigate these issues by deploying the workflow to a computing cluster (for which IDTxl has functionality [39]); the additional compute would allow for the use of a non-linear estimator and the expansion of the search window to several hours.

By procedural limitations, we refer to those that stem from the design of mTE algorithm itself as it is currently implemented within the IDTxl package [39] and the way in which we deployed it here. These include (i) the requirement for equally sampled processes and (ii) lack of discovery of the linear direction of influence. With respect to (i), the requirement to re-sample toxicity processes requires that difficult assumptions be made about the causal effects of communications (see: Sect 4.1). Though this is a current limitation of mTE implementations, future work may negate this requirement and allow for a more natural analysis of point-process like social media data [e.g., 31]. Regarding (ii), at present NTA does not return any information about the direction of influence. This means that it lacks the capacity to identify entities that tend to specifically produce *greater* toxicity in others through their

interactions with them; information that would be of considerable theoretical and practical value. To address this limitation, a refinement of the NTA approach could be to estimate the directionality of effects by integrating a supplementary analysis of the sign of the time-series relationships between entities. By combining the transfer entropy results with a targeted correlation or regression analysis, we could discern the valence of these interactions-distinguishing between the presence of positive and negative influences. This enhancement would enable us to pinpoint entities that notably exacerbate toxicity in others. Future research should address these limitations and explore the broader applicability of NTA to other social media platforms. Overall, NTA provides a promising approach to understanding and mitigating toxic behaviours online, paving the way for targeted interventions and strategies to foster healthier and more constructive digital discourse.

Appendix A: Chat/channel classification codebook

Chat/channel topical classifications were preserved (as below) from the original dataset, see: [32].

A.1 Categories:

- **pol** (in general, such as elections)
- **right** (right-wing politics)
- **left** (left-wing politics)
- **free** (activism for freedom of speech)
- **farm** (farmers protest groups)
- **act19** (activism against COVID-19 measures)
- **riot** (riots in response to COVID-19 measures)
- **news** (alternative news)
- **anti-est** (explicitly anti-establishment groups)
- **c19** COVID conspiracy (groups dedicated to discussing corona as a conspiracy)
- **con** (groups where spirituality and conspiracy thinking fuses together)
- **spirit** (groups dedicated to religious or spiritual activities)

- **health** (health and wellness group)
- **doxx** (public revealing of personal information and attack on individuals)

Appendix B: Toxicity network data

See Fig. 7.

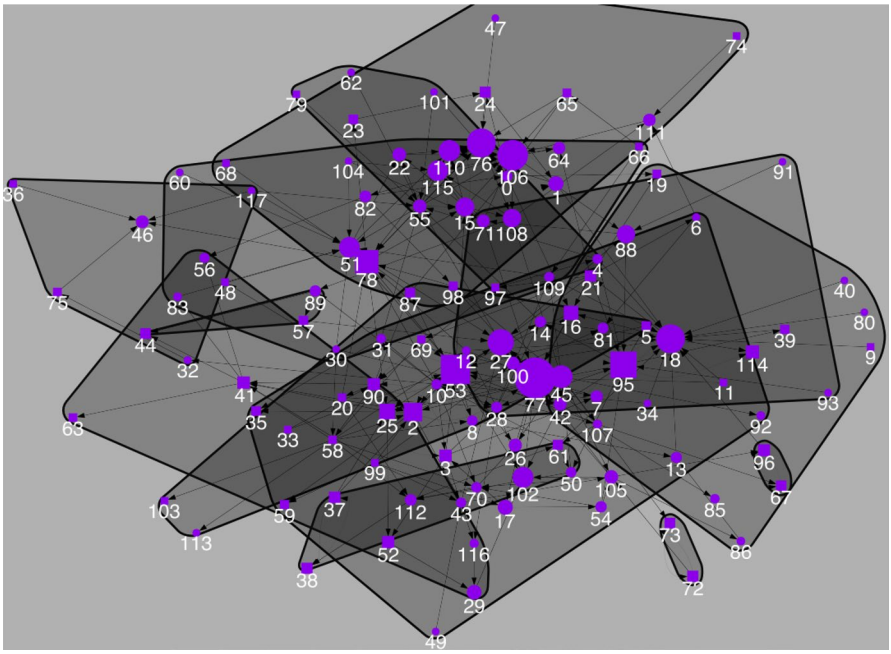

Fig. 7 Structure of toxicity transfer relationships within the Dutch Telegramsphere, January to June 2021. See accompanying table below for details of chats/channels (Table 1). Chat/channel topical classifications were preserved from the original dataset [32]. Influence is measured via PageRank [13]. Toxicity is equal to the proportion of messages with a greater than 50% probability of being toxic

Table 1 Classifications and Measurements of Chats/Channels within the Dutch Telegramsphere (IDs correspond to Fig. 7)

ID	Name	Type	Category	Influence	Toxicity
0	BewustmakendNEWSNOW	Chat	news	0.006843	0.025117
1	DeBataafseRepubliek	Chat	con	0.012989	0.035369
2	DeDagelijkseStandaard	Channel	news	0.018199	0.010195
3	FVDNL	Channel	right	0.008749	0.086314
4	Forumvoordemocratiefvd	Chat	right	0.005215	0.051971
5	FvDMedia	Channel	right	0.004219	0.003152
6	GeweldTegenWakkeren	Chat	free	0.002344	0.014821
7	InfoAvondklok	Chat	act19	0.008648	0.000000
8	JEZUSLEEFTOFFICIAL	Chat	spirit	0.006139	0.006944
9	Juiste	Channel	right	0.001644	0.000000
10	Kletschat_burgers_tegen_onrecht	Chat	free	0.005729	0.000000
11	Liefdespioniers	Channel	spirit	0.001644	0.035306
12	NLPolitiek	Chat	pol	0.003075	0.010817
13	NatuurlijkVrijOnderwijsGO	Chat	free	0.006950	0.033946
14	Nederland_in_het_verzet	Chat	free	0.007115	0.000000
15	QPatriotsEindhoven	Chat	con	0.018633	0.004768
16	RedPillJournal	Channel	con	0.012028	0.018049
17	SpoedWet	Chat	act19	0.013062	0.000000
18	VechtVoorRechtt	Chat	right	0.033905	0.022472
19	Waarheid_Eindtijd_Profetie	Channel	con	0.003507	0.037491
20	Wakker2020	Channel	news	0.002711	0.000000
21	WakkerWezen	Channel	news	0.007408	0.034286
22	WijZijnDeVrijheid	Chat	free	0.010953	0.002381
23	artsenvoorvrijheid	Channel	act19	0.004749	0.013333
24	bataafsenieuws	Channel	news	0.006844	0.026555
25	bewustzijncentrumapofyliet	Channel	spirit	0.013707	0.055556
26	bezorgdeoudersschijndelchat	Chat	act19	0.010149	0.082173
27	boereninopstand	Chat	farm	0.029899	0.155660
28	boereninopstand2_0	Chat	farm	0.006714	0.067731
29	boete_avondklok	Chat	act19	0.012469	0.028986
30	boetemondkapje	Chat	c19	0.001644	0.007315
31	bronvanonvoorwaardelijkeliefde	Chat	spirit	0.004341	0.000000
32	burgerwachtNio	Chat	free	0.002131	0.008475
33	c0r0naboelsjit	Channel	con19	0.001644	0.018054
34	cafeweltschmerz	Chat	news	0.001644	0.039194
35	complotmemes	Channel	con	0.005084	0.012759
36	complotwappies	Channel	con	0.001644	0.018182
37	coronavaccinsbijwerkingen	Channel	con19	0.007687	0.019802
38	coronavaccinschatgroep	Channel	con19	0.007109	0.004246
39	covfefereport	Channel	con	0.004219	0.002597
40	covidwaarheid	Chat	c19	0.001644	0.000000

Table 1 (continued)

ID	Name	Type	Category	Influence	Toxicity
41	dagelijksestandaard	Channel	news	0.009058	0.034998
42	deelwatjeweet	Chat	c19	0.009731	0.000000
43	defendrotterdam	Chat	free	0.005485	0.000000
44	degeboden	Channel	spirit	0.006269	0.000000
45	denhaaginopstand	Chat	act19	0.024771	0.035714
46	deparallellemaatschappij	Chat	free	0.009550	0.007538
47	deredacteur	Chat	news	0.001644	0.070866
48	donmaartenofficial	Channel	news	0.001994	0.000000
49	dvo_open_koffiecorner	Chat	act19	0.001644	0.000000
50	eindhovenrwinachat	Chat	riot	0.005850	0.023256
51	fvd_nl	Chat	right	0.021645	0.144672
52	fvdgeluid	Channel	right	0.008709	0.097012
53	geenstijll	Channel	news	0.034633	0.023319
54	geertwildersss	Chat	right	0.007196	0.020530
55	geloofhoopliefde	Chat	spirit	0.010563	0.022978
56	gewoonvrij	Chat	free	0.005423	0.001369
57	gezelligspiritueel	Channel	spirit	0.004729	0.005692
58	gezondegeest	Channel	spirit	0.003061	0.080000
59	gezondheidineigenhand	Channel	health	0.004087	0.018657
60	ikknaagaandepotenvanRutteenKaag	Chat	free	0.001644	0.000000
61	ikzorgenbenwakker	Channel	con	0.005756	0.028430
62	jongereninopstand	Chat	free	0.001644	0.000000
63	joumij	Channel	spirit	0.003184	0.006098
64	klokkenchat	Chat	free	0.008600	0.024417
65	klokkenluiders	Channel	free	0.003507	0.023014
66	klokkenvideos	Channel	free	0.001644	0.027681
67	kritischezorgverleners	Channel	act19	0.005978	0.028736
68	langefranspodcast	Channel	news	0.001644	0.030062
69	leefbewust	Channel	health	0.002633	0.029799
70	leefbewustnederland	Chat	health	0.005987	0.035104
71	mondkapjesverzetgroep	Chat	c19	0.009662	0.005626
72	natuurlijf	Channel	health	0.007098	0.000000
73	natuurlijf_chat	Channel	health	0.006416	0.025117
74	natuurlijkegezondheid	Channel	health	0.001644	0.035369
75	nederlandinopstand	Channel	anti_est	0.003456	0.004082
76	nederlandsverzet21	Chat	con	0.034096	0.086314
77	neetegen15 m	Chat	act19	0.052631	0.007574
78	onafhankelijkepers	Channel	news	0.025564	0.046939
79	onderzoekenvancomplotten	Channel	con	0.001644	0.039627
80	opstand19	Chat	act19	0.001644	0.000000
81	ouders0497	Chat	act19	0.006605	0.000000
82	oudersmetzorgen	Chat	act19	0.007900	0.006944

Table 1 (continued)

ID	Name	Type	Category	Influence	Toxicity
83	partijvoordevrijheid	Chat	right	0.003184	0.000000
84	platteaarde	Channel	con	0.001644	0.036592
85	rellen023	Chat	riot	0.004402	0.000000
86	rellennl2	Chat	riot	0.003172	0.004082
87	robertjensenshow	Channel	news	0.005320	0.086314
88	samen1voornl	Chat	free	0.017769	0.051971
89	sameninactie	Chat	free	0.007257	0.039627
90	thepostonline	Channel	news	0.008784	0.003152
91	vaccinatiewaarheid	Chat	c19	0.001644	0.000000
92	verkiezingen2021	Chat	pol	0.003399	0.006944
93	vervoervoordemonstranten	Chat	free	0.001644	0.000000
94	vervoervoordemonstrantenchat	Chat	free	0.003042	0.000000
95	verzetsblaadje	Channel	anti_est	0.030289	0.000000
96	virussenbestaanniet2	Channel	con	0.008254	0.010817
97	viruswaanzin	Channel	con19	0.002344	0.033946
98	viruswaarheid_zooms	Channel	con19	0.003883	0.000000
99	vizieroplinks	Channel	doxx	0.001644	0.000000
100	vriendenvangeertwilders	Chat	right	0.010291	0.037553
101	vrijewinkel_nederland	Chat	free	0.001644	0.004768
102	vrijheid	Chat	free	0.021875	0.010821
103	vrouwenvoorvrijheid	Channel	free	0.001933	0.000000
104	waarheidbovenalles	Chat	con	0.001644	0.022472
105	wakeyupp	Chat	con	0.010785	0.035036
106	wakkerAmsterdam	Chat	con	0.036800	0.037491
107	wakkere	Chat	act19	0.003886	0.000000
108	wakkerenchat	Chat	con	0.018091	0.034286
109	wakkereouders	Chat	act19	0.005239	0.002381
110	wakkergroningen	Chat	con	0.022609	0.045455
111	whereWeGo1WeGoAlll	Chat	con	0.009164	0.013333
112	wilderspvv	Chats	right	0.007732	0.006405
113	worldwidenetherlands	Chat	free	0.001933	0.029368
114	wrwynl	Channel	con	0.009983	0.000000
115	wwncommunity	Chat	con	0.021380	0.020408
116	zoektocht	Chat	con	0.004025	0.026555
117	zondermondkapje	Chat	c19	0.001644	0.055556

Declarations

Funding This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 947695).

Conflict of interest The authors of this manuscript declare no competing interests.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and materials The data that support the findings of this study are available from the corresponding author upon request and will be made publicly available before the end of the completion of the overarching research project (ERC-2020-STG-947695).

Code availability the code with which the analyses in this paper were performed is available from the corresponding author upon request, and in addition, a respository containing code (in Python) for the general application of NTA to social media data is publicly available [14].

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Rupert Kiddle. The first draft of the manuscript was written by Rupert Kiddle and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Almerekhi, H., Kwak, H., & Jansen, J., et al. (2019). Detecting toxicity triggers in online discussions. In: HT '19: Proceedings of the 30th ACM Conference on Hypertext and Social Media. ACM Digital Library, pp 291–292, https://doi.org/10.1145/3342220.3344933.
2. Bauer, T.L., Colbaugh, R., & Glass, K., et al. (2013). Use of transfer entropy to infer relationships from behavior. In: Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop on—CSIIRW '13. ACM Digital Library, https://doi.org/10.1145/2459976.2460016.
3. Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review, 116*(1), 1–18. https://doi.org/10.1017/S0003055421000885
4. Bossomaier, T., Barnett, L., Harré, M., et al. (2016). *An Introduction to Transfer Entropy*. Springer International Publishing. https://doi.org/10.1007/978-3-319-43222-9
5. Can, U., & Alatas, B. (2019). A new direction in social network analysis: Online social network analysis problems and applications. *Physica A: Statistical Mechanics and its Applications*. https://doi.org/10.1016/j.physa.2019.122372
6. Cavallaro, L., Ficara, A., Meo, P. D., et al. (2020). Disrupting resilient criminal networks through data analysis: The case of Sicilian Mafia. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0236476
7. Centola, D. (2018). *How Behavior Spreads: The Science of Complex Contagions*. Princeton University Press.
8. Cheng, J., Bernstein, M., & Danescu-Niculescu-Mizil, C., et al. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In: Proceedings of the 2017 ACM

Conference on Computer Supported Cooperative Work and Social Computing. ACM Digital Library, https://doi.org/10.1145/2998181.2998213.
9. Faes, L., Nollo, G., & Porta, A. (2011). Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Physical Review E*. https://doi.org/10.1103/PhysRevE.83.051112
10. Ficara, A., Cavallaro, L., Curreri, F., et al. (2021). Criminal networks analysis in missing data scenarios through graph distances. *PLoS One*. https://doi.org/10.1371/journal.pone.0255067
11. Google, Jigsaw. (2023). Perspective API. https://www.perspectiveapi.com/.
12. Guhl, J., & Davey, J. (2020). A Safe Space to Hate: white Supremacist Mobilisation on Telegram. Tech. rep., Institute for Strategic Dialogue, https://www.isdglobal.org/isd-publications/a-safe-space-to-hate-white-supremacist-mobilisation-on-telegram/.
13. Haveliwala, T., & others. (1999). Efficient computation of PageRank. Tech. rep., Stanford University, https://citeseerx.ist.psu.edu/document?repid=rep1 &type=pdf &doi=83997ceff5acd718cd22c5efab4fe2264938676c.
14. Kiddle, R. (2023). Rptkiddle/NetToxAnalysis. https://github.com/Rptkiddle/NetToxAnalysis.
15. Kim, J. W., Guess, A., Nyhan, B., et al. (2021). The distorting prism of social media: how self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication, 71*(6), 922–946. https://doi.org/10.1093/joc/jqab034
16. Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E, 69*(6), 066138. https://doi.org/10.1103/PhysRevE.69.066138
17. La Morgia, M., Mei, A., & Mongardini, A.M., et al. (2021). Uncovering the dark side of Telegram: fakes, clones, scams, and conspiracy movements. http://arxiv.org/abs/2111.13530.
18. Leeuwen, M.v. (2021). Politie haalt Telegram-kanalen complotdenkers offline om bedreigingen. ADnl https://www.ad.nl/binnenland/politie-haalt-telegram-kanalen-complotdenkers-offline-om-bedreigingen~ac1fe23f/, section: Binnenland.
19. Lizier, J., Rubinov, M. (2012). Multivariate construction of effective computational networks from observational data. https://www.semanticscholar.org/paper/Multivariate-construction-of-effective-networks-Lizier-Rubinov/984ccd9b344b9ec3e7e10672027b57e4a2a4432d.
20. Lizier, J. T., & Prokopenko, M. (2010). Differentiating information transfer and causal effect. *The European Physical Journal B, 73*(4), 605–615. https://doi.org/10.1140/epjb/e2010-00034-5
21. Maleki, M., Arani, M., & Mead, E., et al. (2022). Applying an Epidemiological Model to Evaluate the Propagation of Toxicity related to COVID-19 on Twitter. In: Proceedings of the 55th Hawaii International Conference on System Sciences, https://doi.org/10.24251/HICSS.2022.401.
22. Notarmuzi, D., Castellano, C., Flammini, A., et al. (2022). Universality, criticality and complexity of information propagation in social media. *Nature Communications, 13*(1), 1308. https://doi.org/10.1038/s41467-022-28964-8
23. Obadimu, A., Khaund, T., Mead, E., et al. (2021). Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube. *Information Processing & Management, 58*(5), 102660. https://doi.org/10.1016/j.ipm.2021.102660
24. Pond, T., Magsarjav, S., South, T., et al. (2020). Complex contagion features without social reinforcement in a model of social information flow. *Entropy, 22*(3), 265. https://doi.org/10.3390/e22030265
25. Powers, E., Koliska, M., & Guha, P. (2019). Shouting matches and echo chambers: perceived identity threats and political self-censorship on social media. International Journal of Communication https://www.semanticscholar.org/paper/%E2%80%9CShouting-Matches-and-Echo-Chambers%E2%80%9D%3A-Perceived-and-Powers-Koliska/3e3c878fd938028e37fae5a394f24c5c8a343756.
26. Quandt, T. (2018). Dark participation. *Media and Communication, 6*(4), 36–48. https://doi.org/10.17645/mac.v6i4.1519
27. Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication, 35*(3), 213–229. https://doi.org/10.1177/0267323120922066
28. Salminen, J., Sengün, S., Corporan, J., et al. (2020). Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLoS One*. https://doi.org/10.1371/journal.pone.0228723
29. Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters, 85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

30. Mn, Serrano, Boguñá, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences, 106*(16), 6483–6488. https://doi.org/10.1073/pnas.0808904106
31. Shorten, D. P., Spinney, R. E., & Lizier, J. T. (2021). Estimating transfer entropy in continuous time between neural spike trains or other event-based data. *PLOS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1008054
32. Simon, M., Welbers, K., Kroon, C. A., et al. (2023). Linked in the dark: A network approach to understanding information flows within the Dutch Telegramsphere. *Information, Communication & Society, 26*(15), 3054–3078. https://doi.org/10.1080/1369118X.2022.2133549
33. Steeg, G.V., & Galstyan, A. (2012). Information transfer in social media. In: Proceedings of the 21st international conference on World Wide Web. ACM Digital Library, pp 509–518, https://doi.org/10.1145/2187836.2187906.
34. Steeg, G.V., & Galstyan, A. (2013). Information-theoretic measures of influence based on content dynamics. http://arxiv.org/abs/1208.4475.
35. Sunstein, C. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press, https://press.princeton.edu/books/hardcover/9780691175515/republic.
36. Trilling, D. (2022). Beyond echo chambers and filter bubbles: Towards a feedback-loop model of political communication. Prague, Czech Republic, https://newsflows.eu/wp-content/uploads/2022/06/epsa2022.pdf.
37. Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS One*. https://doi.org/10.1371/journal.pone.0203958
38. Vogels, E., Anderson, M., & Nolan, H., et al. (2021). The State of Online Harassment. Tech. rep., Pew Research Center, https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.
39. Wollstadt, P., Lizier, J., Vicente, R., et al. (2019). IDTxl: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software, 4*(34), 1081. https://doi.org/10.21105/joss.01081
40. Zeng, J., & Schäfer, M. S. (2021). Conceptualizing dark platforms. COVID-19-related conspiracy theories on 8kun and gab. *Digital Journalism, 9*(9), 1321–1343. https://doi.org/10.1080/21670811.2021.1938165