PLOS ONE

RESEARCH ARTICLE

Structural invariants and semantic fingerprints in the "ego network" of words

Kilian Ollivier*, Chiara Boldrini, Andrea Passarella, Marco Conti

CNR-IIT, Pisa, Italy

* kilian.ollivier@iit.cnr.it

OPEN ACCESS

Citation: Ollivier K, Boldrini C, Passarella A, Conti M (2022) Structural invariants and semantic fingerprints in the "ego network" of words. PLoS ONE 17(11): e0277182. https://doi.org/10.1371/journal.pone.0277182

Editor: Diego Raphael Amancio, University of Sao Paulo, BRAZIL

Received: February 1, 2022

Accepted: October 21, 2022

Published: November 22, 2022

Data Availability Statement: All relevant data files are available from the OSF database (osf.io/gmpaz).

Funding: This work was partially funded by the SoBigData++, HumaneAI-Net, and SAI projects. The SoBigData++ project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042. The HumaneAI-Net project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952026. The SAI project

Abstract

Well-established cognitive models coming from anthropology have shown that, due to the cognitive constraints that limit our "bandwidth" for social interactions, humans organize their social relations according to a regular structure. In this work, we postulate that similar regularities can be found in other cognitive processes, such as those involving language production. In order to investigate this claim, we analyse a dataset containing tweets of a heterogeneous group of Twitter users (regular users and professional writers). Leveraging a methodology similar to the one used to uncover the well-established social cognitive constraints, we find regularities at both the structural and semantic levels. In the former, we find that a concentric layered structure (which we call *ego network of words*, in analogy to the ego network of social relationships) very well captures how individuals organise the words they use. The size of the layers in this structure regularly grows (approximately 2-3 times with respect to the previous one) when moving outwards, and the two penultimate external layers consistently account for approximately 60% and 30% of the used words, irrespective of the number of layers of the user. For the semantic analysis, each ring of each ego network is described by a semantic profile, which captures the topics associated with the words in the ring. We find that ring #1 has a special role in the model. It is semantically the most dissimilar and the most diverse among the rings. We also show that the topics that are important in the innermost ring also have the characteristic of being predominant in each of the other rings, as well as in the entire ego network. In this respect, ring #1 can be seen as the semantic fingerprint of the ego network of words.

1 Introduction

In humans, language production is a deliberate and conscious action. However, it relies on many invisible mental processes that allow the construction of sentences in a very short time. For example, these cognitive processes are at play during the word retrieval stage, when the brain has to efficiently process, in a few milliseconds, its lexicon in order to find the right word, among thousands of others, that best fits the concept that needs to be expressed [1]. In order to achieve this impressive performance, cognitive strategies that exploit language properties, such as word frequency (e.g. when the most frequently used words are retrieved more

is supported by the CHIST-ERA grant CHIST-ERA-19-XAI-010, by MUR (grant No. not yet available), FWF (grant No. I 5205), EPSRC (grant No. EP/V055712/1), NCN (grant No. 2020/02/Y/ST6/00064), ETAg (grant No. SLTAT21096), BNSF (grant No. КП-06-ДОО2/5). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

quickly [2, 3]), are activated. In this paper, we set out to find traces of these cognitive patterns in written production with a data-driven approach. To this end, we rely on the ego network model, which has already uncovered the cognitive limits of another human activity: socialisation.

1.1 The social ego network model

Anthropologists have shown that the number of meaningful social relationships that humans can maintain is not only limited to 150 [4] (the famous Dunbar's number) but it is also stable over time. The discovery of this regularity in human activity stems from the observation that, in different species of primates, there exists a correlation between the size of the neocortex (the part of the brain dedicated to high-level cognitive functions such as socialisation, language, etc.), and the average size of groups in natural environments. Extrapolating the expected size of a human group from the dimension of the human brain, as well as studying historical data such as the maximum size before fission of autonomous communities [5], the Dunbar number consistently emerges. It was then shown that these 150 active social relationships can be further subdivided into 4 *concentric* circles [6, 7], the innermost one containing the most intimate social relationships [8], the outermost one enclosing all 150 social relationships. The typical size of these concentric circles is 5, 15, 50, and 150, respectively, with a constant scaling ratio of about 3 between consecutive circles. Note that the portion of a circle not included in its inner-most ones is referred to as *ring*. This hierarchical structure of social relationships is called "ego network". Recent studies based on data collected from online social networks have shown that online relationships are subject to the same laws as offline ones: the size of the ego network (i.e., the total number of social relationships) remains in the same order of magnitude as the Dunbar's number, which indicates that the cognitive constraint yielding this number is not overridden by a communication medium that facilitates social interactions [8–11]. In OSNs (Online Social Networks), the typical number of circles is slightly higher than 4, due to the presence of an additional circle in the center of the ego network (containing about 1.5 people), but the scaling ratio is preserved at around 3 (Fig 1).

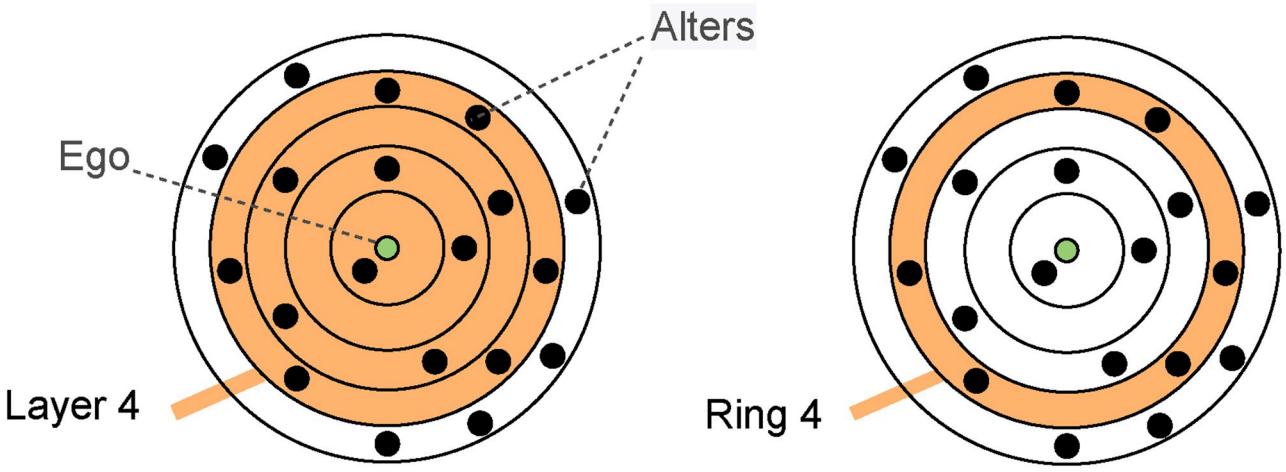

Fig 1. The *ego network* of social relationships. The green dot symbolizes the ego and the black dots the alters with whom the ego maintains an active social relationship. A layer also contains the alters of the inner layers, unlike the rings.

https://doi.org/10.1371/journal.pone.0277182.g001

1.2 From social ego networks to ego networks of words

The ego network model highlights the regularity of the structure of social relations, in real life and in OSN. In this paper, we adopt an analogous approach to investigate the regularities and invariants manifesting cognitive constraints in language production. Specifically, we conjecture that a similar structure, which we call *"ego network of words"*, may also be used to describe the way humans use words, and that this structure may provide very significant information to characterise the peculiarities of individuals, similarly to the social dimension. In fact, it is known [12] that many traits of social behavior (resource sharing, collaboration, diffusion of information) are chiefly determined by the structural properties of social ego networks.

The motivation for this analogy is twofold. First, the use of words is, much like socialisation, a process that involves the use of cognitive resources, thus we conjecture that the ego network model may have larger applicability in describing how humans allocate cognitive resources, for example to language. Second, language is a social activity, whose emergence is potentially linked to the surge in active human relationships from the 50 of the closest primate to 150 for humans. This theory, known as *social gossip theory of language evolution* [13], postulates that language facilitates grooming social relations by reaching several peers at the same time. In addition, there is already well-established knowledge of a number of empirical cognitive limits affecting language, such as the bounded size of our vocabulary (which is consistently limited to approximately 42, 000 words for a native 20-year-old English speaker [14]), as well as the Zipf's law of words [15], which states that the frequency of a word is inversely proportional to its position in the frequency table for most human writings. We, therefore, choose to study the individual distribution of vocabulary, by forming concentric circles of words according to their frequency of use by the ego in question. Then, going beyond words as units of language, we focus on the topics to which the words refer. We thus complement the structural analysis with a semantic study, which completes our cognitive analysis framework. In the same way that the social ego network model has been used to provide a different perspective to social network analysis (such as for information diffusion [16]), we want to leverage the ego networks of words as microscopes to discover novel properties of language production.

1.3 Contribution and key findings

The main contribution of this work is the structural and semantic analysis of the ego networks of words for Twitter users. By using the ego network model, in this paper, we uncover complex structures showing that the cognitive effort to organise one's vocabulary is limited in many ways. We choose a corpus of text made up of tweets because it allows us to work with a varied sample of "authors" (e.g. more varied than a corpus of newspaper articles). Moreover, as Twitter is dedicated to the exchange of very short messages (240 characters), it is a medium that is very favourable to spontaneous reactions, with a more natural style and a reduced writing time. This time constraint is more likely to reveal human behaviour, in analogy with the social domain, where time limitations have been shown to significantly affect social cognitive constraints [13]. For our data-driven analysis, we collected tweets from generic as well as specialised Twitter users (Section 3). Using the ego-network-of-words model, we are able to find evidence of a structural regularity in the frequency of word usage by each individual (Section 4). The semantic analysis (Section 5) also establishes the existence of additional invariants, but most importantly it uncovers the nature of the innermost layer as the *semantic fingerprint* of the whole ego network, i.e., this layer groups together the most important topics on which the user is active. This strengthens the analogy with the social version of the ego network model, where the innermost layers include the most important social relationships of a person.

The key findings of the paper are the following.

- Similarly to the social case, we found that a *regular concentric, layered structure* (which we call *ego network of words* in analogy to the ego networks of the social domain) very well captures how an individual organizes their cognitive effort in language production. Specifically, words can be typically grouped in between 5 and 7 layers of decreasing usage frequency moving outwards, regardless of the specific class of users (regular vs professional).
- One structural invariant is observed for the *size of the layers*, which approximately doubles when moving from layer i to layer $i+1$. The only exception is the innermost layer, which tends to be approximately 5 five times smaller than the next one. This suggests that the innermost layer, the one containing the most used words, may be drastically different from the others.
- A second structural invariant emerges for the *external layers*. Users with more layers organise differently their innermost layers, without modifying significantly the size of the most external ones. In fact, while the size of all layers beyond the first one linearly increases with the most external layer size, the second-last and third-last layers consistently account for approximately 60% and 30% of the used words, irrespective of the number of layers of the user.
- The semantic analysis of the words contained in the ego networks confirms that layer #1 is exceptional in the ego networks of words: it generates proportionally more topics than the other rings, these topics are more diverse, and its overall semantic profile is the most different with respect to those of other rings.
- In addition, topics that are important in ring #1 tend to be important in other rings as well (we call this the *pulling power* of ring #1). Thus, layer #1, despite being the smallest, can be seen as the *semantic fingerprint* of the ego network of words.
- The topics that are primary in some rings tend to be stronger than average among the primary and non-primary topics in the semantic profile of the other rings. This shows that, while layer #1 provides a particularly strong signal about prevalence in the ego networks, weaker signals show a more complex structure of influence among topics "resident" in different layers of the ego network of words.

This paper extends our prior publication in [17], where the structural analysis was carried out. Specifically, in this paper, we also present an extensive semantic analysis of the ego network of words. This allows us to provide a much more comprehensive understanding of the model, and highlight ways to characterise specificities of individuals as they emerge from their use of words, in addition to structural invariants observed through the structural properties of the ego networks.

2 Related work

To the best of our knowledge, no work has been published yet on models of individual word organisation similar in spirit to ours (i.e., by exploring the analogy with the social ego network model). However, some work has already been done on individual word frequency distribution by extending the notion of Zipf's law [18]. Based on Zipf's law, some have tried to find a generative model that could explain such a regularity-based human cognition [19], or just how the limited capacities of our memory naturally constrain our long-term use of words [20]. More generally, vocabulary size is often studied in the context of language learning for both children and adults, as well as to detect possible cognitive impairments [21]. For the semantic part, we have not identified any previous work on modelling user interests with a stratified

approach, such as ours, that relies on the ego network of words. Most publications are about topic recommendations (relying upon a wide range of techniques, such as hashtag analysis [22], LDA [23] or ontology databases [24]), and about the emergence and monitoring of trending topics on Twitter [25, 26].

3 The dataset

The analysis is built upon four datasets extracted from Twitter, using the official Search and Streaming APIs (note that the number of downloadable tweets—at the time of download—was limited to the most recent 3200 tweets per user). Each of them is based on the tweets issued by users in four distinct groups:

Journalists Extracted from a Twitter list containing New York Times journalists (https://twitter.com/i/lists/54340435), created by the New York Times itself. It includes 678 accounts, whose timelines have been downloaded on February 16th, 2018.

Science writers Extracted from a Twitter list created by Jennifer Frazer (https://twitter.com/i/lists/52528869), a science writer at *Scientific American*. The group is composed of 497 accounts and has been downloaded on June 20th, 2018.

Random users #1 This group has been collected by sampling among the accounts that posted a tweet or a retweet in English with the hashtag *#MondayMotivation* (at the download time, on January 16th, 2020). This hashtag is chosen in order to obtain a diversified sample of users: it is broadly used and does not refer to a specific event or a political issue. This group contains 5183 accounts after bot filtering.

Random users #2 This group has been collected by sampling among the accounts that posted a tweet or a retweet in English, from the United Kingdom (we set up a filter based on the language and country), at download time on February 11th, 2020. This group contains 2733 accounts after bot removal.

These four groups are chosen to cover different types of users: the first two contain accounts that use language professionally (journalists and science writers) and the other two contain regular users, which are expected to be more colloquial and less controlled in the language they use. Since the random user accounts are not handpicked as in the two first groups, we need to make sure that they represent real humans. The probability that an account is a bot is calculated with the Botometer service [27], which implements a state-of-the-art bot detection algorithm. This probability that the account is not human, which is called "complete automation probability" (CAP), is not only based on linguistic features such as grammatical tags, or the number of words in a tweet, but also on language-agnostic features like the number of followers or the tweeting frequency [28]. There is no standard CAP threshold to easily separate bots from humans: it depends on the expected balance of precision and recall. That is why we discard accounts with a CAP higher than 0.5, which considerably limits the number of false negatives (undetected bots). The Botometer service achieves a performance of 0.95 AUC on standard bot detection datasets [27]. With this configuration, the algorithm detects 29% of bot accounts in the dataset of random users#1 and 23% in the dataset of random users#2.

In our analysis, we only consider the timelines of *active* Twitter accounts, i.e., users that tweet regularly. Since this preprocessing step largely follows the standard approach in the related literature [8, 29], further details are left to the S1 Appendix. Please note that we discard retweets with no associated comments, as they do not include any text written by the target user, and tweets written in a language other than English (since most of the NLP tools needed for our analysis are optimised for the English language).

3.1 Extracting user timelines with the same observation period

As discussed above, for each user in our datasets we retrieved the most recent 3200 tweets (due to the Twitter API limitation), which constitute the *observed timeline* of the user. The time period covered by these tweets varies according to the frequency with which the account is tweeting: for very active users, the last 3200 tweets will only cover a short time span. Since random users are generally more active, their observation period is shorter, and this may create a significant sampling bias. In fact, the length of the observation period affects the measured word usage frequencies (specifically, we cannot observe frequencies lower than the inverse of the observation period). In order to guarantee a fair comparison across user categories and to be able to compare users with different tweeting activities without introducing biases, we choose to work on timelines with the same duration, by restricting to an observation window T. To obtain timelines that have the same observation window T (in years), we delete all those with a duration shorter than T and remove tweets written more than T years ago from the remaining ones.

Increasing T reduces the number of users we can keep for our analysis (see Fig 2): for a T larger than 2 years, that number is halved, and for a T larger than 3 years, it falls below 500 for all datasets. On the contrary, the average number of tweets per timeline increases linearly with T (Fig 3). The choice of an observation window will then result from a trade-off between a high number of timelines per dataset and a large average number of tweets per timeline. To simplify the choice of T, we only select round numbers of years. We can read in Table 1 that, beyond 3 years, the number of users falls below 100 for some datasets. On the other hand, the number of tweets for $T = 1$ year remains acceptable (> 500). Since we value the diversity of users (in order to limit any bias in the selection of Twitter accounts) over the number of tweets available, we make the choice of $T = 1$ year for the entire paper. Results with other T lengths can be found in [17]. We note that random users have a higher frequency of tweeting than

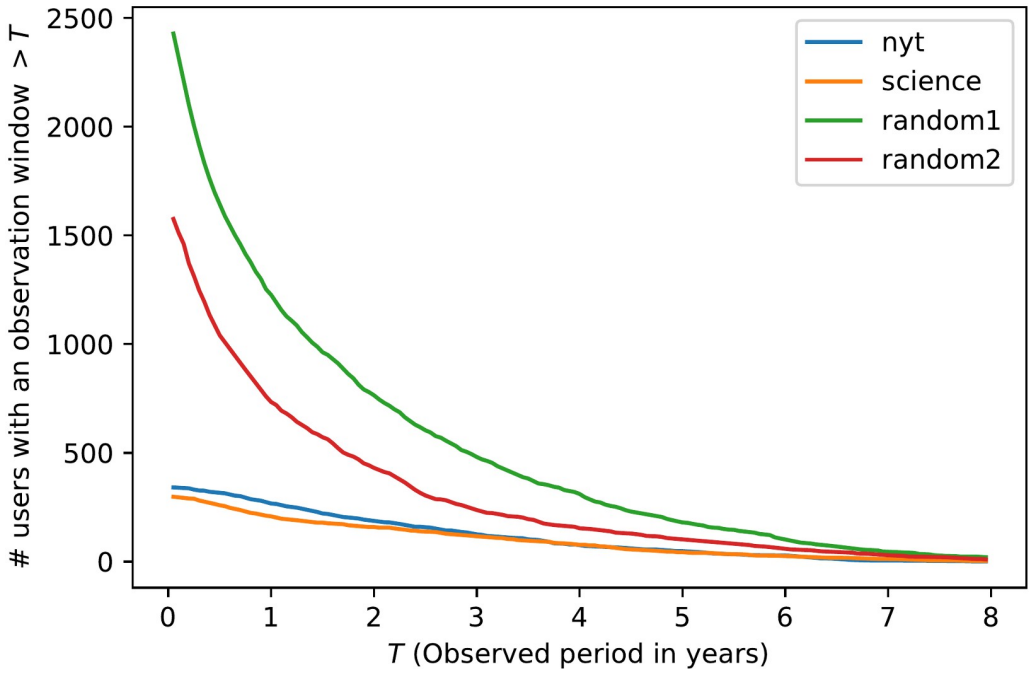

Fig 2. Available timelines. Number of selected timelines depending on the observation window.

https://doi.org/10.1371/journal.pone.0277182.g002

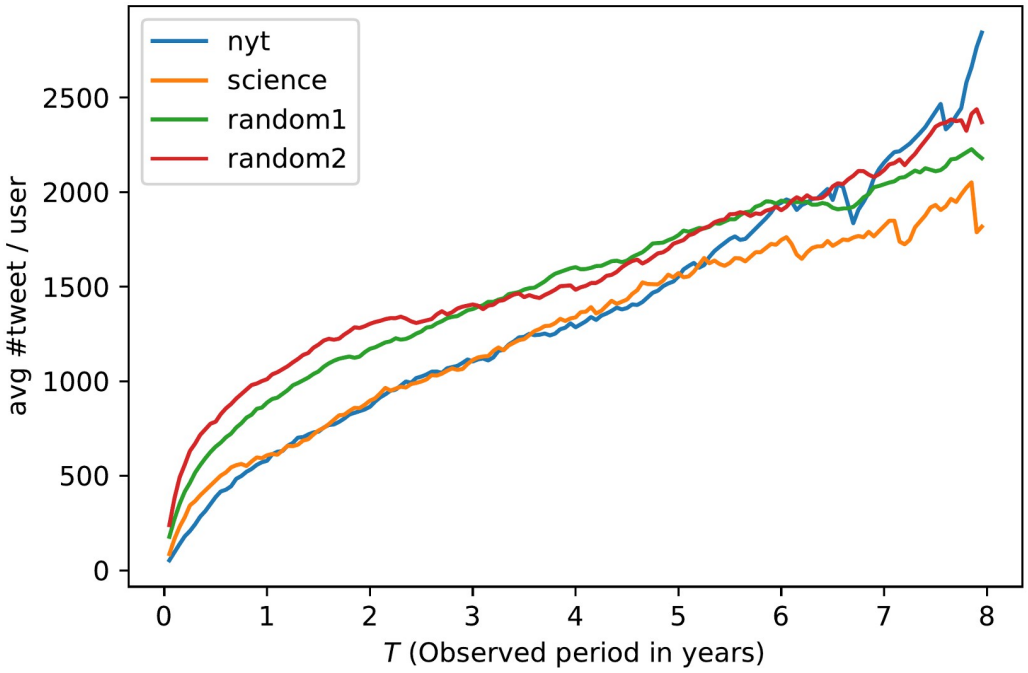

Fig 3. Tweets per user. Average number of tweets depending on the observation window. The Pearson linear correlation coefficient is equal to or greater than .98 for the four datasets.

https://doi.org/10.1371/journal.pone.0277182.g003

others. This difference tends to smooth out when the observation period is longer (Table 1). This can be explained by the fact that the timelines with the highest tweeting frequency are excluded in that case because their observation period is too small (which further supports the fact that a smaller T reduces the selection bias of users).

4 Structural analysis of the ego network of words

In this section, we focus on the analysis of structural properties of the ego network of words, highlighting structural invariants in language production. Note that, in the social domain, pure structural properties of ego networks were instrumental [12] in characterising many traits of social behavior (resource sharing, collaboration, diffusion of information). For this reason, we believe it is important to assess them in the language domain as well, before moving on (Section 5) to more complex and domain-specific analyses.

We first describe the methodology we use for our analysis in Section 4.1, then we discuss the results in Section 4.2. For ease of reading, the notation used in this section is summarised in Table 2. The section reports only the most significant results obtained by analysing the

Table 1. Datasets summary. Number of users and tweeting frequency at different observation windows.

Datasets	Number of users			Avg # of tweets / user		
	1 year	2 years	3 years	1 year	2 years	3 years
NYT Journalists	268	187	125	579.71	865.02	1104.58
Science Writers	208	159	117	609.08	897.29	1112.63
Random Users #1	1227	765	311	897.29	1179.98	1403.50
Random Users #2	734	431	153	1057.41	1315.71	1404.60

https://doi.org/10.1371/journal.pone.0277182.t001

Table 2. Summary of notation used in the structural analysis.

Name	Notation	Definition/formula
Optimal number of circles	$\tau^{(e)}$	the results of the clustering on the word frequencies for the user (ego) e
Circle (or layer)	$\mathcal{L}_i^{(e)}$	i-th social circles of the tagged ego e, with $i \in \{1, \ldots, \tau^{(e)}\}$
Scaling ratio of layer i	$\rho_i^{(e)}$	$\frac{\lvert\mathcal{L}_i^{(e)}\rvert}{\lvert\mathcal{L}_{i-1}^{(e)}\rvert}$, with $i \in \{2, \ldots, \tau^{(e)}\}$
Ring	$r_i^{(e)}$	$\mathcal{L}_i^{(e)} - \mathcal{L}_{i-1}^{(e)}$

https://doi.org/10.1371/journal.pone.0277182.t002

structural properties of the ego network. Interested readers are referred to [17] for additional results.

4.1 Methods

For each user, acting as ego, we want to build their ego network of words. To this aim, we first extract individual words from the user's tweets (Section 4.1.1), then we build the actual ego network from these words (Section 4.1.2).

4.1.1 Word extraction. Since the analysis focus on words and their frequency of use, we take advantage of NLP techniques for extracting them. As a first step, all the syntactic marks that are specific to communication in online social networks (mentions with @, hashtags with #, links, emojis) are discarded (see S1 Appendix for a summary). Once the remaining words are tokenized (i.e., identified as words), those that are used to articulate the sentence (e.g., "with", "a", "but") are dropped. In linguistics, this type of word is called a functional word as opposed to lexical words, which have a meaning independent of the context. These two categories involve different cognitive processes (syntactic for functional words and semantic for lexical words), different parts of the brain [30], and probably different neurological organizations [31]. We are more interested in lexical words because their frequency in written production depends on the author's intentions, as opposed to functional word frequencies that depend on language characteristics. Functional words may also depend on the style of an author (and due to this they are often used in stylometry). Still, whether their usage requires a significant cognitive effort is arguable, hence in this work, we opted for their removal. Moreover, lexical words represent the biggest part of the vocabulary. Functional words are generally called stop-words in the NLP domain and we simply used an already existing list from the library spaCy [32] to remove them.

As this work will leverage word frequencies as a proxy for discovering cognitive properties, we need to group words derived from the same root (e.g. "work" and "worked") in order to calculate their number of occurrences. This operation can be achieved with two methods: stemming and lemmatization. Stemming algorithms generally remove the last letters thanks to complex heuristics, whereas lemmatization uses the dictionary and a real morphological analysis of the word to find its normalized form. Stemming is faster, but it may cause some mistakes in overstemming and understemming. For this reason, we choose to perform lemmatization with the help of the package WordNetLemmatizer from the library NLTK [33] (which leverages the lexical database WordNet). Once we have obtained the number of occurrences for each word base, we remove all those that appear only once to leave out the majority of misspelled words. The S1 Appendix contains examples of the entire preprocessing part.

In the remaining of the paper, when we talk about the "words" of a user, we refer to the set of words left after removing functional words and after lemmatization.

4.1.2 Building the ego network of words. Let us focus on a user j. When studying the social cognitive constraints [28], the contact frequency between two people was taken as a

proxy for their intimacy and, as a result, for their cognitive effort in nurturing the relationship. Similarly, the frequency f_i at which user j uses word i is considered here as a proxy of their "relationship". Frequency f_i is given by $\frac{n_{ij}}{T}$, where n_{ij} denotes the number of occurrences of word i in user j's timeline, and T denotes the observation window of j's timeline in years ($T = 1y$ in our case, as discussed in Section 3.1). Using this frequency definition, we now investigate whether the words of a user can be grouped into homogeneous classes and whether different users feature a similar number and sizes of classes. To this aim, for each user, we leverage a clustering algorithm to group words with a similar frequency. The selected algorithm is Mean Shift [34], because as opposed to Jenks [35] or k-means [36], it is able to automatically detect the optimal number of clusters. In order to account for the long-tailed nature of frequencies, a standard log-transformation is applied to the frequency values prior to the Mean Shift run.

Thus, for each user, we feed the user's words to Mean Shift. The output of the clustering process is one value $\tau^{(e)}$ for each ego network e, which describes the optimal number of classes (clusters) in which the word frequencies can be split. We rank each cluster by its position in the frequency distribution: cluster #1 is the one that contains the most frequent words, and the last cluster is the one that contains the least used words. Following the convention of the social ego network model discussed in Section 1, these clusters can be mapped into concentric layers (or circles), which provide a cumulative view of word usage. Specifically, layer $\mathcal{L}_i$ includes all clusters from the first to the i-th. Layers provide a convenient grouping of words used *at least* at a certain frequency. We refer to this layered structure as the *ego network of words*. Note that, since layers in ego networks are cumulative (i.e., they include all words used at least a certain frequency), we will use the term "ring" to refer to their non-overlapping portion: for example, ring #2 contains all words that are in $\mathcal{L}_2$ but not in $\mathcal{L}_1$ (see Table 4 for the general formula). For the sake of example, let us focus on the second cluster identified by Mean Shift: cluster #2 corresponds to ring #2 in the ego network, and the union of ring #1 and ring #2 corresponds to the 2nd layer of the ego network. Another typical metric that is analysed in the context of social cognitive constraints is the scaling ratio ρ_i between layers i and $i - 1$, which, as discussed earlier, corresponds to the ratio between the size of consecutive layers (see Table 4 for its formula). The scaling ratio is an important measure of regularity, as it captures a relative pattern across layers, beyond the absolute values of their size. Taken together, the optimal number of layers $\tau^{(e)}$, the circle $\mathcal{L}_i^{(e)}$, and the scaling ratio $\rho_i^{(e)}$ fully characterise the ego network e.

4.2 Results

Here we study the ego networks of words in our four datasets, following the methodology described above.

The histograms of the obtained optimal number of layers τ are shown in Fig 4. It is interesting to note that, despite the heterogeneity of users (in terms of tweeting frequency), the distributions are always quite narrow, with peaks appearing consistently between 5 and 7 clusters. Similarly to the social constraints case, also for language production, we observe a fairly regular and consistent structure. This is the first important result of the paper, hinting at the existence of structural invariants in cognitive processes.

We now study the size of the layers identified in Fig 4. For the sake of statistical reliability, we only consider those users whose optimal number of layers (as identified by Mean Shift) corresponds to the most popular number of layers (red bars) in Fig 4. This allows us to have a sufficient number of samples in each class. Fig 5 shows the average layer sizes for every dataset. For a given number of clusters, we observe again a striking regularity across the datasets, meaning that each layer has approximately the same size regardless of the category of users.

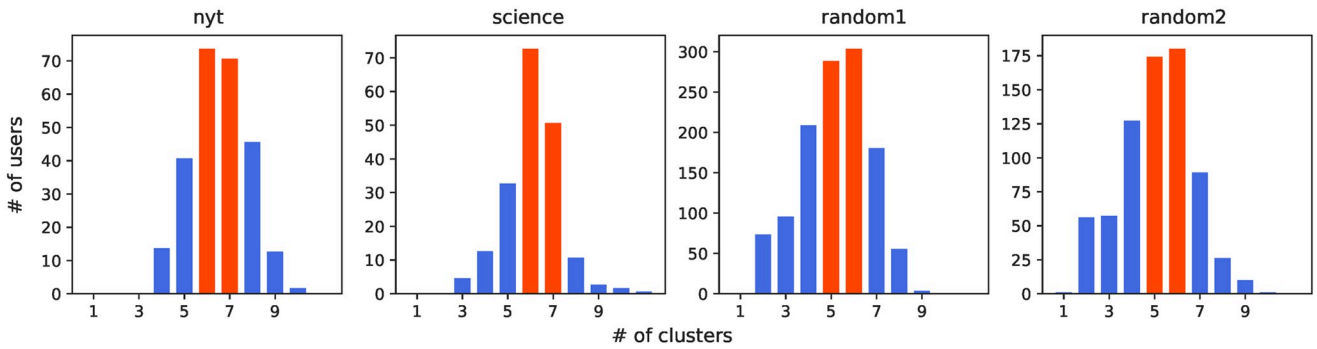

Fig 4. Optimal number of clusters. The clusters are obtained by applying Mean Shift to log-transformed frequencies. The most frequent number of clusters is highlighted in red.

https://doi.org/10.1371/journal.pone.0277182.g004

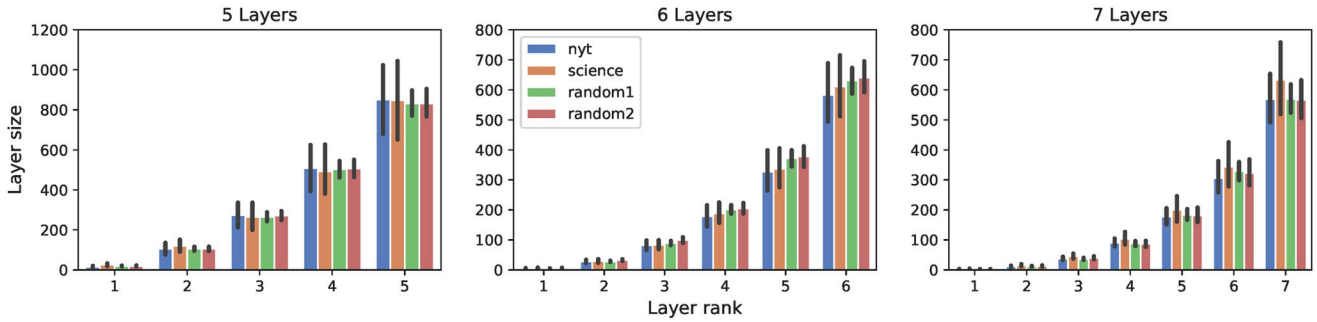

Fig 5. Average layer size. Each panel captures egos with a different optimal number of clusters. Error bars correspond to the 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0277182.g005

Fig 6 shows the scaling ratio of the layers in language production. We can observe the following general behavior: the scaling ratio starts with a high value between layers #1 and #2, but always gets closer to 2–3 as we move outwards. This empirical rule is valid whatever the dataset (and whatever the observation period [17]). This is another significant structural regularity, quite similar to the one found for social ego networks, as a further hint of cognitive constraints behind the way humans organise the words they use.

In order to further investigate the structure of the word clusters, we compute the linear regression coefficients between the total number of unique words used by each user

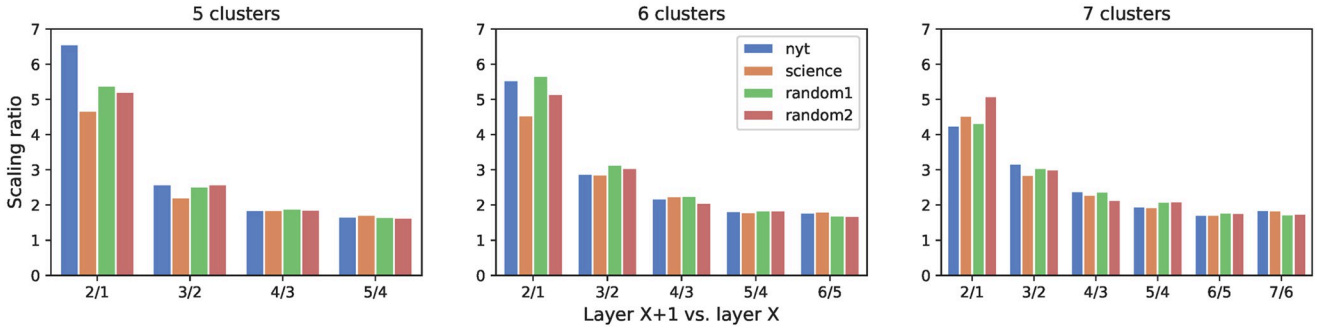

Fig 6. Scaling ratio. Each panel captures egos with a different optimal number of clusters. Error bars correspond to the 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0277182.g006

Table 3. Size of external layer vs individual layer size: Regression coefficients. We report the linear regression coefficients obtained for the journalists dataset with $T = 1$ year.

Opt. # of clusters	Cluster Rank						
	1	2	3	4	5	6	7
5 clusters	0.02	0.13	0.33	0.62	1.00		
6 clusters	0.01	0.04	0.14	0.32	0.59	1.00	
7 clusters	0.00	0.02	0.06	0.16	0.32	0.56	1.00

https://doi.org/10.1371/journal.pone.0277182.t003

(corresponding to the size of the outermost layer) and the individual layer sizes. Due to space limits, in Table 3 we only report the exact coefficients for the journalists' dataset (but analogous results are obtained for the other categories) and in Fig 7 we plot the linear regression for all the user categories. Note that the size of the most external cluster is basically the total number of words used by an individual in the observation window. It is thus interesting to see what happens when this number increases, i.e., if users who use more words distribute them uniformly across the clusters, or not. Table 3 shows two interesting features. First, it shows another regularity, as the size of all layers linearly increases with the most external cluster size, with the exception of the first one (Fig 7). Moreover, it is quite interesting to observe that the second-last and third-last layers consistently account for approximately 60% and 30% of the used words, irrespective of the number of clusters. This indicates that users with more clusters split, at a finer granularity, words used at the highest frequencies, i.e., they organise differently their innermost clusters, without modifying significantly the size of the most external ones.

As a final comment on Fig 6, please note that the innermost layer tends to be approximately five times smaller than the next one. This suggests that this layer, containing the most used words, may be drastically different from the others (as also evident from Table 3). The characterization of this special layer will be the main focus of the next section.

4.3 Discussion

We summarise below the main results of the section.

- Individual distributions of word frequencies are divided into a consistent number of groups. Since word frequencies impact the cognitive processes underlying word learning and retrieval in the mental lexicon [37], these groups can be an indirect trace of these processes' properties. The number of groups is only marginally affected by the class (specialized or generic) the users belong.

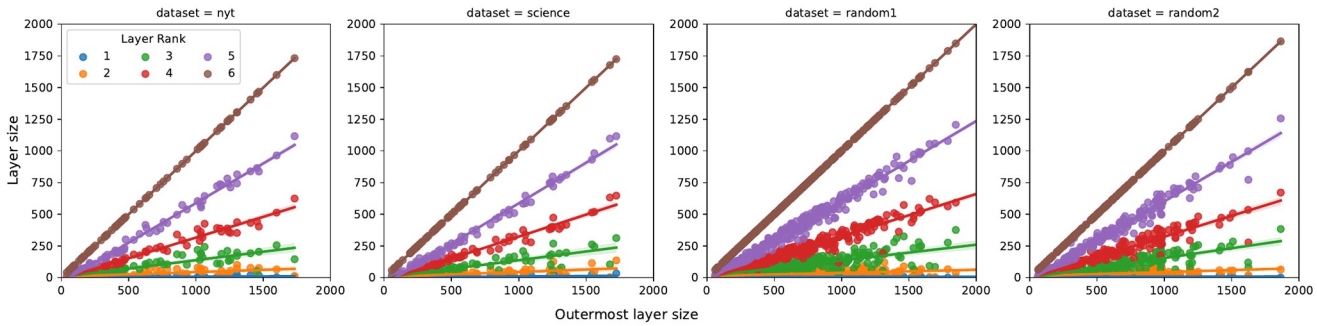

Fig 7. Size of external layer vs individual layer size: Linear regression plots. The x-axis corresponds to the total number of unique words used by each user (corresponding to the size of the outermost layer), the y-axis to the individual layer sizes.

https://doi.org/10.1371/journal.pone.0277182.g007

- Structural invariants in terms of layer sizes and scaling ratio are observed, similarly to the well-known results from the social domain [8]. Specifically, we found that the size of the layers approximately doubles when moving from layer i to layer $i + 1$, with the only exception of the first layer.
- Users with more layers organise differently their innermost layer, without modifying significantly the size of the most external ones, which consistently account for approximately 60% and 30% of the used words, irrespective of the number of clusters of the user.

5 Semantic analysis of the ego network of words

We have treated words as simple tokens so far. However, words have meanings and they can be linked to specific topics. In this section, we want to go beyond words and investigate which topics they refer to and how they are distributed in the different rings of the ego network. The analysis of this section revolves around the concept of *semantic profile* of a ring (in the ego network of words), which captures the topics associated with the words in the ring. Once semantic profiles are obtained, we are able to address the following high-level question: are all rings similar in the topics they contain, or does the ego network organize the topics in its rings in a specific way?

For the convenience of the reader, we summarise in Table 4 the notation used throughout the section.

5.1 How to build semantic profiles

In this section, we describe how we carry out the semantic analysis of the ego network of words. First, in Section 5.1.1, we motivate our selection of the BERTopic framework for topic extraction. Then, in Section 5.1.2, we illustrate the steps for topic extraction. At the end of this process, each word occurrence in the ego network is associated with a specific topic. Accounting for the popularity of each topic in the rings of the ego network, in Section 5.1.3 we build the *semantic profile* of the ego network ring, as the topic distribution of the words in that ring.

5.1.1 Preliminaries. To calculate a semantic profile, we choose to consider the meaning of each word in its context rather than using a semantic dictionary [38] (a dataset where each word is mapped to a semantic category), which would not be able to detect more complex topics and would miss some meanings for a polysemous word. We acknowledge that a lot of effort has been put in the direction of ontologies in order to understand more precisely the interests of users, specifically on Twitter. Ontologies map knowledge of specific domains, such as Athena [24], which is a semantic web database extracted from a news portal that can be used for news recommendation purposes [39], or the BBC ontologies extracted from the BBC corpus of news, which allows politically-oriented topic mining [40]. However, even if their drawbacks (such as the rigidity of the knowledge model) can be partly fixed by coupling them with models based on embedding [41], we prefer having the maximum freedom in the topic identification process by using a transformers-based model such as BERT [42] which is the current state of the art in text embedding and then using an unsupervised method to detect topics.

5.1.2 Extraction of the topics. In order to avoid some issues with polysemous words, we must consider the ring of an ego network not only as a set of single words associated with a frequency of use but as a set of words with a given number of occurrences (from which the frequency is derived), each occurrence belonging to a user's tweet. We aim to associate each word occurrence with a topic. We first classify (in an unsupervised way) the tweets by topic using the BERTopic framework [43], then all word occurrences that constitute a tweet are assigned the same topic as the tweet itself (Fig 8).

Table 4. Summary of the notation used in the semantic analysis.

Symbol	Description
$\mathcal{E}$	Set of all ego networks $\mathcal{E}$
$e \in \mathcal{E}$	Ego networks e belonging to the set of all ego networks $\mathcal{E}$
$c \in \mathcal{C}$	Topic c belonging to the set of all topics $\mathcal{C}$
$m \in \mathcal{T}$	Tweet m belonging to the set of all tweets $\mathcal{T}$
P_m	Semantic profile of tweet m, according to HDBSCAN
$P_m(c)$	Likelihood that tweet m belongs to topic c, according to the semantic profile of the tweet
$\mathcal{W}(e, r)$	Set of non-distinct words in ring r of ego network e
$\mathcal{W}_u(e, r)$	Set of distinct words in ring r of ego network e
$\mathcal{W}(e, w_u)$	Set of occurrences of the unique word w_u in the ego network e
$O(e, r)$	Number of word occurrences in ring r of ego network e
$o(w_u, e)$	Number of occurrences associated with the unique word w_u of ego network e
$P_r^{(e)}$	Semantic profile of ring r of ego network e
$P_r^{(e)}(c)$	Probability of observing topic c in $P_r^{(e)}$ of ring r in ego network e
$P_{w_u}^{(e)}$	Topic distribution of unique word w_u in ego network e
$P_{w_u}^{(e)}(c)$	Probability of observing topic c in $P_{w_u}^{(e)}$ for w_u in ego network e
$\mathcal{N}(e, r)$	The number of topics discussed in ring r of ego network e
$\mathcal{N}_{norm}(e, r)$	$\mathcal{N}(e, r)$ normalised by the total number of word occurrences in r
$H(e, r)$	Entropy of the semantic profile $P_r^{(e)}$
$\delta_{JS}(P_{r_i}^{(e)} \| \| P_{r_j}^{(e)})$	Distance between the semantic profiles of rings i and j
$U_r^{(e)}$	Set of primary topics for ring r of ego network e
$L_r^{(e)}$	Set of non-primary topics for ring r of ego network e
$K_{TOP(r_x)}^{r_y}$	Coverage of r_x's primary topics in r_y's semantic profile
$S_{TOP(r_x)}^{r_y}$	Strength of r_x's primary topics in r_y's semantic profile
$S_{BOTTOM}^{r_y}$	Strength of r_x's non-primary topics in r_y's semantic profile
$S_{TOP(r_x, r_y)}^{r_y}$	Strength of topics that are primary for both r_x and r_y in r_y's semantic profile
$S_{TOP(r_x), BOTTOM(r_y)}^{r_y}$	Strength of topics that are primary for r_x but not for r_y in r_y's semantic profile
$\sigma_{TOP(r_x, r_y)}^{r_y}$	Strength of topics that are primary for both r_x and r_y with respect to the average strength of primary topics in r_y's semantic profile
$\sigma_{TOP(r_x), BOTTOM(r_y)}^{r_y}$	Strength of topics that are primary for r_x but not for r_y with respect to the average strength of non-primary topics in r_y's semantic profile

https://doi.org/10.1371/journal.pone.0277182.t004

For the current analysis, we chose to focus only on ego networks with six rings, the case covering the most users. As described in the following, the BERTopic framework uses sequentially BERT [42] for tweet embedding, UMAP [44] for dimension reduction, and HDBSCAN [45] for clustering those tweet embeddings in a low-dimensional subspace.

5.1.2.1 Tweet embedding with BERT. BERT [42], which achieves state-of-the-art performance for natural language understanding, is used to assign to each tweet a point in the embedding space which is supposed to be a vector representation of its semantic meaning. BERT is a bidirectional transformer developed by Google, trained on the BookCorpus [46] and Wikipedia in English. It, therefore, relies on all the linguistic knowledge learned from a very large corpus to perform this task. BERT yields topics along 768 dimensions.

5.1.2.2 Dimensionality reduction with UMAP. In order to mitigate the curse of dimensionality (to which clustering algorithm based on k-nearest neighbors are particularly sensible [47]), we use the UMAP clustering algorithm (with settings `n neighbors = 15, n components = 5, metric='cosine'` and the python package `umap v0.1.1`) to reduce the

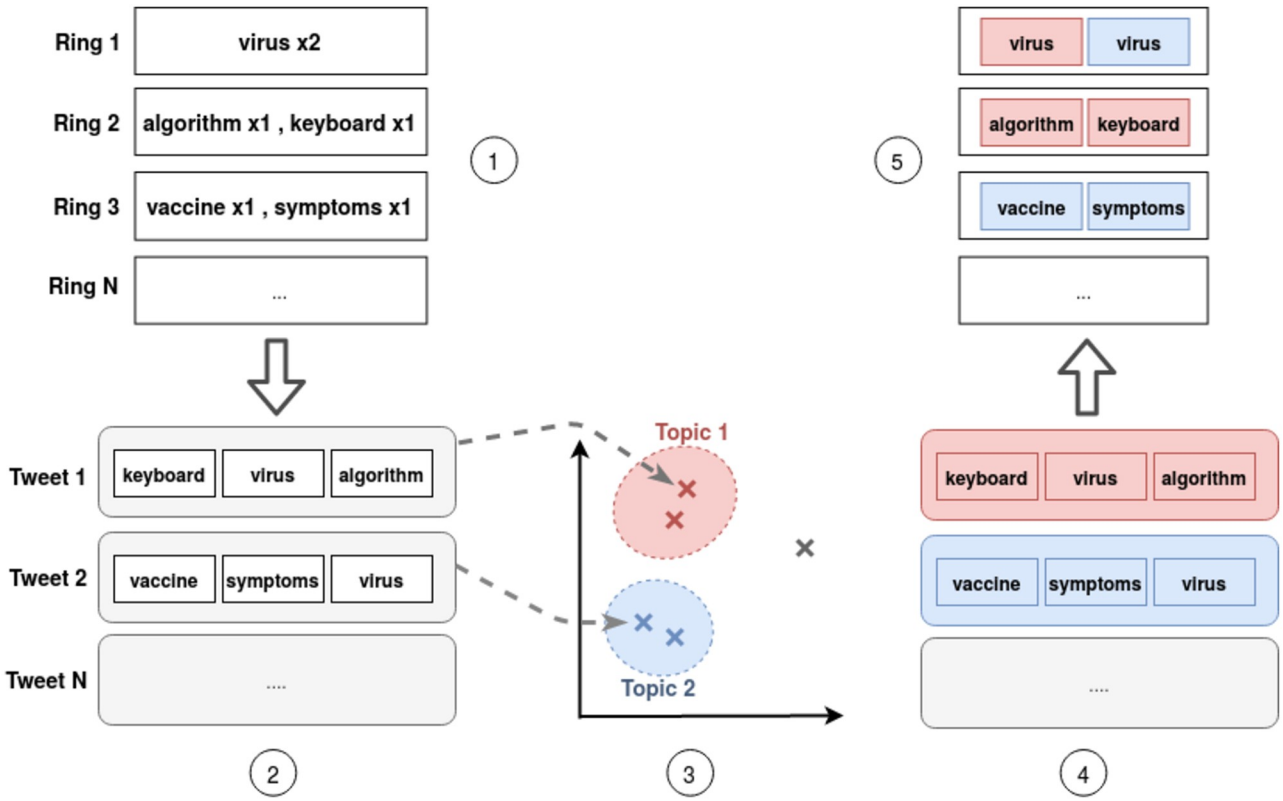

Fig 8. Obtaining the semantic profile of the rings of an ego network. (1) The ego network's rings organize a user's vocabulary based on the frequencies of the words. (2) For a given word, its occurrences in the user timeline are coming most likely from different tweets. (3) The tweets are classified by topic thanks to the BERTopic framework. (4) Each word occurrence is assigned the very same topic as the tweets it belongs to. (5) If we consider a ring as a multiset of words (with repetitions) the semantic profile is the distribution of the topics among those words.

https://doi.org/10.1371/journal.pone.0277182.g008

embedding space down to five dimensions as recommended in the BERTopic framework [43]. UMAP, like the T-SNE [48] algorithm, is able to capture latent non-linear dimensions but in a more scalable way.

5.1.2.3 HDBSCAN for clustering topics. HDBSCAN [45] is also able to find non-linear cluster structures from the density, as well as outliers, like DBSCAN (Fig 9). However, instead of deciding the contours of a cluster based on a fixed density threshold, HDBSCAN uses hierarchical clustering (single linkage) to find the most stable partition. Here we use HDBSCAN with following settings: `min cluster size = 15, metric='euclidean', cluster selection method='eom', prediction data = True` with the python package `hdbscan v0.8.26`. Thanks to BERT embedding, the clusters of tweets we obtain are semantically homogeneous, and therefore represent the dominant topics of the dataset. Under these conditions, we can consider that a cluster corresponds to a topic.

Table 5 shows the percentage of outliers detected by HDBSCAN, which corresponds to the percentage of tweets that cannot be associated with a specific topic. Since this percentage is quite high, even with the most conservative configurations (with the least outliers), we also assess the cluster configuration (*i.e.*, the topic assignment) induced by a soft clustering approach. Indeed HDBSCAN allows two types of clustering: hard clustering, which classifies each tweet in one and only one cluster (or as an outlier), and soft clustering, which is able to measure the proximity of a tweet to several different clusters. The advantage is that it is possible to obtain this proximity even for outliers, which allows us to integrate them into the

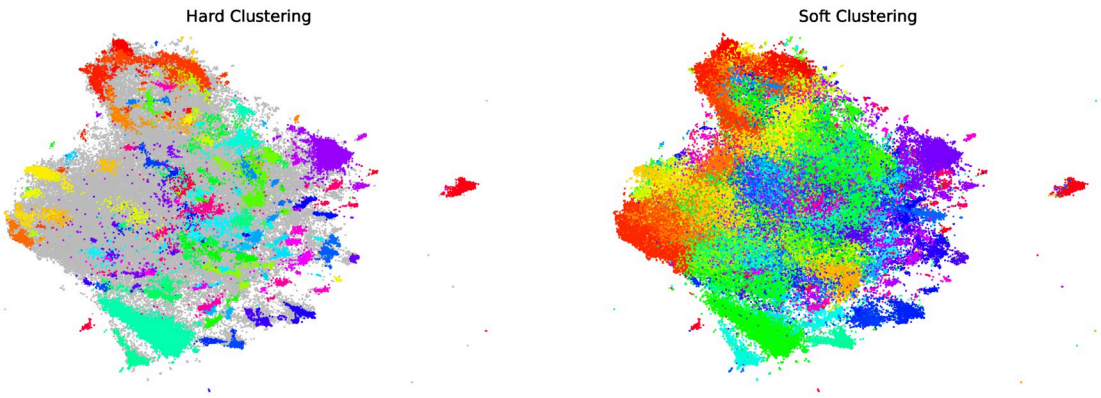

Fig 9. 2D visualization of the HDBSCAN results on the journalists dataset with both hard and soft clustering. 265 clusters are found (they are the same in both cases). In the first case, each point is classified as either belonging to a single cluster (colored points) or as an outlier (grey point), whereas in the second case each point is assigned a likelihood to belong to each cluster (the points take the color of the cluster they belong to most likely).

https://doi.org/10.1371/journal.pone.0277182.g009

analysis. When using it for soft clustering, HDBSCAN provides, for each point (tweet) m, a probability distribution P_m such that $P_m(c)$ is the likelihood that this point belongs to the cluster (topic) c, with $\sum_{c\in\mathcal{C}} P_m(c) \leq 1$ ($\mathcal{C}$ being the set of topics). Thus, with soft clustering, the tweet is not assigned a single topic but a probability distribution over all the topics. For clarity reasons, in the case of hard clustering—where the tweet m is directly assigned one topic c_m—let us use the same notation P_m, where $P_m(c_m)$ is equal to 1 and zero otherwise. We will use these two configurations (hard clustering and soft clustering) to build two separate semantic profiles for each ego network ring. In S1 Appendix we discuss in detail why hard clustering is better suited for our analysis.

5.1.2.4 Reduction of the number of topics. As shown in Table 5, the different datasets feature a different number of topics. In order to be able to compare the datasets, we reduced the number of topics down to the same number of topics (this set of topics—which is different for each dataset—will be noted as $\mathcal{C}$ from now on). Let us denote with $\mathcal{C}'$ the full set of topics. Our goal is to merge them together until we obtain the target number of topics. To do so, the following operation is repeated: merge the smallest cluster c'_1 (in the hard clustered configuration) with the cluster c'_2 to which c'_1 is semantically the closest. This semantic similarity is calculated as follows: all the tweets are grouped in a single document by cluster, then a TF-IDF vector is calculated for each of them. The similarity between the two topics is the cosine of their TF-IDF representation. The probability of the new topic $c'_1 \cup c'_2$ is accordingly updated, for each tweet m, as $P_m(c'_1 \cup c'_2) = P_m(c'_1) + P_m(c'_2)$. When merging step by step the clusters, the average similarity between them increases as can be seen in Fig 10. In the case of journalists and science writers, we see that exceeding 100 topics no longer allows the emergence of topics that are radically different from the others, while still enabling an acceptable number of topics to be

Table 5. Topics per dataset. Each topic corresponds to a cluster identified by HDBSCAN.

Datasets	Number of topics	% of outliers
NYT Journalists	265	69.3%
Science Writers	223	71.8%
Random Users #1	2940	68.6%
Random Users #2	2577	70.0%

https://doi.org/10.1371/journal.pone.0277182.t005

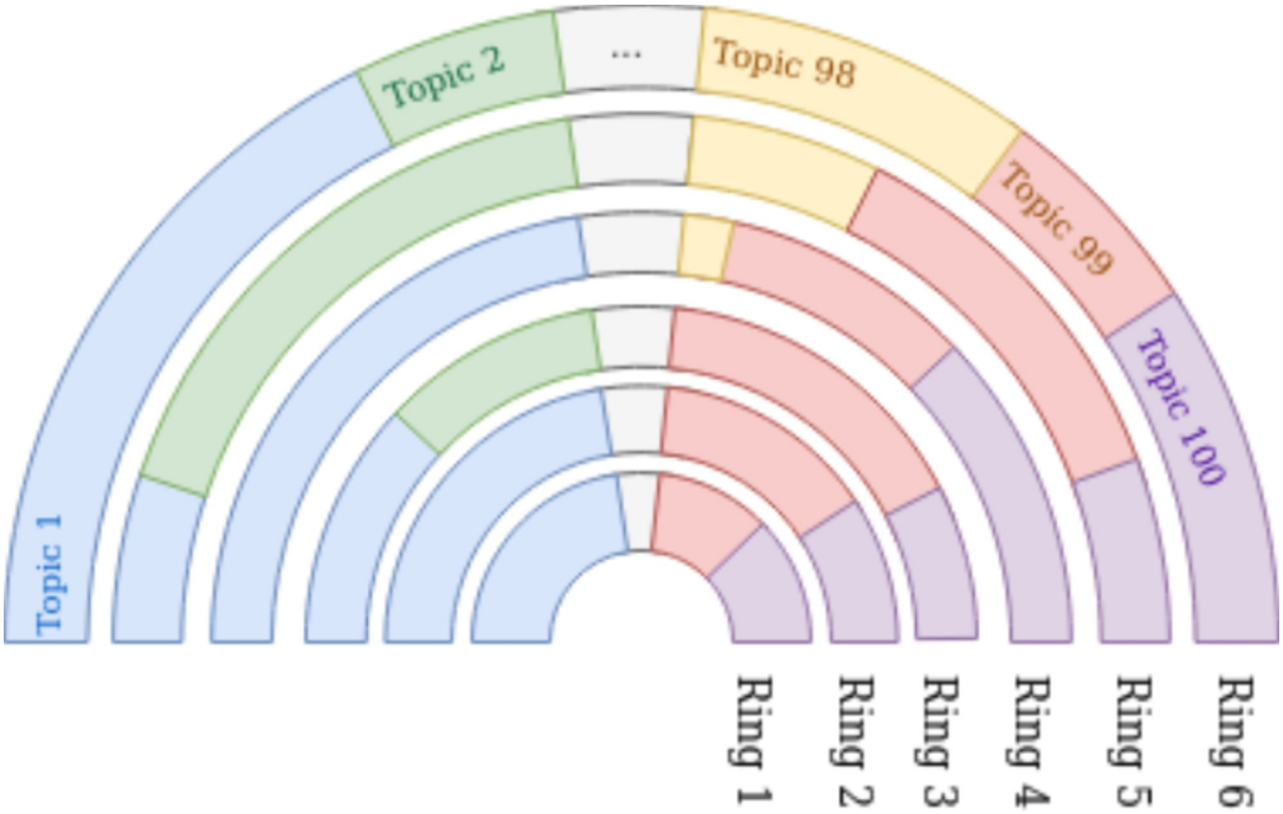

Fig 10. Number of topics vs. average topic similarity. The threshold of one hundred topics is marked with the dashed red line. This threshold is situated at the end of the bend for specialized datasets, and in the middle of the bend for both random datasets.

https://doi.org/10.1371/journal.pone.0277182.g010

isolated. Thus, in order to be able to compare the results related to the different datasets, we have chosen to limit the number of topics to 100 for each of them. For the sake of comparison, the 100 topics obtained for the hard clustering configuration are also used for topic reduction in the soft clustering case. This operation allows us to narrow down to one hundred topics the different semantic fields addressed in the same dataset while trying to provoke the least changes in the topic reassignment.

5.1.3 Extraction of the semantic profile. We define the semantic profile of an ego network ring as the distribution of topics to which the word occurrences that the ring contains (multiple occurrences of the same word may come from different contexts and thus refer to different topics) belong. Note that this analysis is carried out at the ring level, and not the circle level because circles are concentric and cumulative, thus the semantic profiles of circles would include by default overlapping topics, hence creating a bias in the analysis (similarly to counting topics twice). After the preprocessing described in the previous section, each word occurrence is associated with a topic (or several, in the soft clustered case), thus we can compute for each ego network's ring a topic distribution based on the word occurrences it contains.

Let $\mathcal{W}(e, r)$ be the set of word occurrences contained in ring r of the ego network e, and m (w) the tweet the word occurrence w belongs to. The probability $P_r^{(e)}(c)$ of observing topic c in ring r of ego network e is defined as follows:

$$P_r^{(e)}(c) = \frac{\sum_{w \in \mathcal{W}(e,r)} P_{m(w)}(c)}{\sum_{c \in \mathcal{C}} \sum_{w \in \mathcal{W}(e,r)} P_{m(w)}(c)}, \quad (1)$$

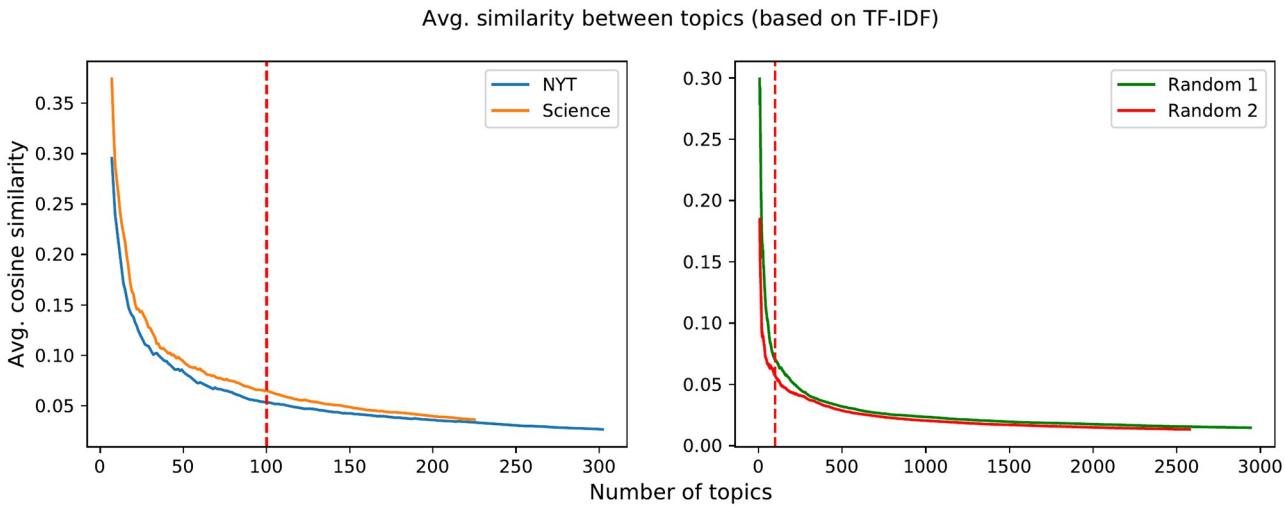

Fig 11. Semantic profile illustration. Each ring is associated with a topic distribution. *Note*: Two different semantic profiles can be built, depending on whether topics are assigned using hard vs soft clustering. In S1 Appendix we show that the use of soft clustering (and thus the inclusion of outliers) does not improve the reliability of the analysis. It gives too much importance to noisy data which favors the emergence of very generalized "super topics" that dominate all semantic profiles. We, therefore, present in Section 5.3 only the results obtained with hard clustering. In S1 Appendix we discuss soft versus hard clustering in detail and motivate why hard clustering is better suited for our analysis.

https://doi.org/10.1371/journal.pone.0277182.g011

where $\sum_{c\in\mathcal{C}} P_r^{(e)}(c) = 1$. More in general, we denote with $P_r^{(e)}$ the semantic profile of ring r in ego network e (depicted in Fig 11). For this reason, we will also refer to $P_r^{(e)}(c)$ as the share of c in the semantic profile $P_r^{(e)}$ of r This unique semantic profile will be the starting point for all subsequent analyses in this section. In S1 Appendix, we provide four tables (one for each dataset) that detail for every topic the most characteristic words and the average share in the rings.

5.2 Metrics for the analysis of semantic profiles

After following the steps described in Section 5.1, we end up with a semantic profile for each ring of an ego network. In the following we discuss (i) how to characterise individual semantic profiles (Section 5.2.1), (ii) how to compare semantic profiles (Section 5.2.2), and (iii) how to leverage semantic profiles to investigate the role of the most important topics (Section 5.2.3).

5.2.1 Characterization of the semantic profile. Let us consider a ring r of ego network e for which we have extracted the semantic profile as discussed above. The semantic profile tells us how many distinct topics the words in ring r touch upon. Formally, the number of topics associated with a given ring can be calculated as follows:

$$\mathcal{N}(e,r) = \sum_{c\in\mathcal{C}} \mathbb{1}_{P_r^{(e)}(c)>0}, \tag{2}$$

where we denoted with $P_r^{(e)}(c)$ the probability of a observing topic c in the semantic profile $P_r^{(e)}$ of ring r, and $\mathbb{1}$ is the indicator function. Note, though, that $\mathcal{N}(e,r)$ may offer only a partial perspective. In fact, rings have very different sizes (as discussed in Section 4) and it is expected to be much easier for larger rings (i.e., rings containing many words) to span a larger range of

topics. For this reason, we will compare $\mathcal{N}(e,r)$ with its normalised version:

$$\mathcal{N}_{norm}(e,r) = \frac{\mathcal{N}(e,r)}{|\mathcal{W}(e,r)|}, \tag{3}$$

where we weigh the number of topics "generated" by the ring by the number of word occurrences contained in the ring (denoted with $|\mathcal{W}(e,r)|$).

$\mathcal{N}(e,r)$ and $\mathcal{N}_{norm}(e,r)$ account for the mere presence of topics, regardless of their frequency of use. To capture the latter dimension, we next measure the entropy of $P_r^{(e)}$. Recalling that $P_r^{(e)}$ is in fact a probability distribution, its Shannon entropy reflects its diversity: the entropy (and diversity) is maximum if a ring contains all topics equally (i.e., with the same values of $P_r^{(e)}(c)$), while the entropy is minimum if a ring contains only one topic. So, the greater the entropy, the greater the diversity. Denoting with $H(e, r)$ the entropy of the ring r in ego e, its definition is as follows:

$$H(e,r) = -\sum_{c\in\mathcal{C}} P_r^{(e)}(c) \times \log\left(P_r^{(e)}(c)\right). \tag{4}$$

For the 100 topics we consider, the minimum entropy is 0 and the maximum entropy is about 4.60.

In Section 5.3, the average of $\mathcal{N}(e,r)$, $\mathcal{N}_{norm}(e,r)$, and $H(e,r)$ across all ego networks will be presented, i.e., $\mathcal{N}(r) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mathcal{N}(e,r)$ (analogously for the others).

5.2.2 Comparing the semantic profiles of different rings. Once we know which topics are covered by each ring of an ego network, the first step is to find out whether their semantic profile differs from one ring to another one or, instead, if the distribution is homogeneous over the whole ego network. Since all semantic profiles are based on the same 100 topics, it is easy to obtain a distance measure to compare the rings with one another. Recalling that the semantic profile is a probability distribution, for this purpose we can use the Jensen-Shannon (JS) divergence [49], which allows us to calculate the proximity between the 100-topic distributions that we obtained previously. Then, the corresponding JS distance is conventionally obtained as the square root of the JS divergence [50]. The JS divergence is basically a symmetric version of the well-known Kullblack-Leibler (KL) divergence, which is a standard metric for capturing the distance between probability distributions. For a tagged ego e, the KL divergence D_{KL} between two semantic profiles $P_{r_i}^{(e)}$ and $P_{r_j}^{(e)}$ of rings i and j for ego network e can be computed as follows:

$$D_{KL}\left(P_{r_i}^{(e)}||P_{r_j}^{(e)}\right) = \sum_{c\in\mathcal{C}} P_{r_i}^{(e)}(c) \times log\left(\frac{P_{r_i}^{(e)}(c)}{P_{r_j}^{(e)}(c)}\right). \tag{5}$$

From $D_{KL}(P_{r_i}^{(e)}||P_{r_j}^{(e)})$, the JS divergence can be obtained as:

$$D_{JS}\left(P_{r_i}^{(e)}||P_{r_j}^{(e)}\right) = \frac{D_{KL}(P_{r_i}^{(e)}||M) + D_{KL}(P_{r_j}^{(e)}||M)}{2}, \tag{6}$$

with $M = \frac{P_{r_i}^{(e)} + P_{r_j}^{(e)}}{2}$. Then we go from divergence D to distance δ by taking the square root: $\delta_{JS}(P_{r_i}^{(e)}, P_{r_j}^{(e)}) = \sqrt{D_{JS}(P_{r_i}^{(e)}||P_{r_j}^{(e)})}$. Note that the JS distance is bounded as $0 \leq \delta_{JS}(P_{r_i}^{(e)}||P_{r_j}^{(e)}) \leq \sqrt{log(2)} \approx 0.83$.

Once we have obtained a $\delta_{JS}(P_{r_i}, P_{r_j})$, we compute its average across all ego networks in a standard way, i.e., $\delta_{JS}^{(e)}\left(P_{r_i}, P_{r_j}\right) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\delta_{JS}^{(e)}\left(P_{r_i}, P_{r_j}\right)$

5.2.3 Capturing important topics and their cross-rings effects. Given a semantic profile $P_r^{(e)}$, we can check whether some topics are more important than others, and, if this is the case, whether they play a special role in the ego network's rings. We consider whether topics can be divided in two classes, *i.e.*, "important" and "not-important" topics for each ring. To do so, we cluster the topics according to their presence in the specific ring under study, i.e, according to the values of $P_r^{(e)}(c)$ where $c \in \mathcal{C}$. To this aim, we use the Jenks algorithm [51] which allows finding natural breaks in the frequency distribution (similarly to k-means, we have to specify k, the number of groups we want to obtain). We rely on the Silhouette score [52] to validate the clustering results. Since we just want to find one natural break that separates important topics from the others, we set $k = 2$. Words are split into two groups, one with high-frequency use, and the other with low-frequency use. The former is the set of important (or primary) topics referred to as $U_r^{(e)}$ (where e is the ego network and r is the ring number), and the latter is the set of non-important topics as $L_r^{(e)}$.

Once we have obtained $U_r^{(e)}$ and $L_r^{(e)}$, for all ego networks and for all rings, we can investigate whether primary topics in one ring play a special role in other rings as well. Let us focus on two rings x and y. We define $K_{TOP(r_x)}^{r_y}$ as the coverage of r_x's primary topics in ring r_y. This metric captures the cumulative presence of r_x's primary topics in r_y.

$$K_{TOP(r_x)}^{r_y} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\sum_{c\in U_{r_x}^{(e)}} P_{r_y}^{(e)}(c). \tag{7}$$

Then, to capture the average individual strength of r_x's primary topics in r_y, we define a complementary metric $S_{TOP(r_x)}^{r_y}$ (with an averaging factor $\frac{1}{|U_{r_x}^{(e)}|}$) as follows:

$$S_{TOP(r_x)}^{r_y} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\frac{1}{|U_{r_x}^{(e)}|}\sum_{c\in U_{r_x}^{(e)}} P_{r_y}^{(e)}(c). \tag{8}$$

Basically, $S_{TOP(r_x)}^{r_y}$ measures the average share of *each* r_x's primary topics in another ring of the same ego network. Similarly, we can compute $S_{BOTTOM(r_x)}^{r_y}$ by replacing $U_{r_x}^{(e)}$ with $L_{r_x}^{(e)}$ in the above equation. This approach can be generalized to more complex cases. For example, we can study the strength of topics that are important in *both* r_x and r_y in the semantic profile of ring r_y. This would be equivalent to the following:

$$S_{TOP(r_x,r_y)}^{r_y} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\frac{1}{|U_{r_x}^{(e)}\cap U_{r_y}^{(e)}|}\sum_{c\in U_{r_x}^{(e)}\cap U_{r_y}^{(e)}} P_{r_y}^{(e)}(c). \tag{9}$$

Analogously, we can study the opposite effect, i.e., what is the strength of topics that are important in r_x but *not* in r_y in the semantic profile of r_y. In this case, the formula will be the following:

$$S_{TOP(r_x),BOTTOM(r_y)}^{r_y} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\frac{1}{|U_{r_x}^{(e)}\cap L_{r_y}^{(e)}|}\sum_{c\in U_{r_x}^{(e)}\cap L_{r_y}^{(e)}} P_{r_y}^{(e)}(c). \tag{10}$$

All the above metrics capture the *pulling power* of ring r_x on ring r_y.

Another interesting perspective is whether topics that are primary elsewhere tend to be more or less dominant than the average topic in $U^{(e)}_{r_y}$ or $L^{(e)}_{r_x}$. This effect can be measured as follows:

$$\sigma^{r_y}_{TOP(r_x,r_y)} = S^{r_y}_{TOP(r_x,r_y)} - S^{r_y}_{TOP(r_y)}, \tag{11}$$

where we basically compute the difference between the strength of topics that are primary in both r_x and r_y and the average strength of all primary topics in r_y. The complementary perspective is whether topics that are primary elsewhere tend to be more or less dominant than the average non-primary topic in r_y. To this aim, we leverage the following:

$$\sigma^{r_y}_{TOP(r_x),BOTTOM(r_y)} = S^{r_y}_{TOP(r_x),BOTTOM(r_y)} - S^{r_y}_{BOTTOM(r_y)}. \tag{12}$$

which follows the same line of reasoning as $\sigma^{r_y}_{TOP(r_x,r_y)}$.

5.3 Results

In this section, we study the semantic profiles in the ego networks of the Twitter users in our four datasets (Section 3).

5.3.1 Ring #1 is special in the ego networks of words. We start our analysis by studying how topics are associated with the different rings. For each ego network e, we will compute the number of topics per ring ($\mathcal{N}(e, r)$ and $\mathcal{N}_{norm}(e, r)$, its normalized version) and their entropy $H(e, r)$. These metrics are then averaged across all egos, as described in Section 5.3, and 95% confidence intervals are shown.

In Fig 12(a), we can observe that the number of topics grows towards the external rings (from about 11 in ring #1 to over 16 in ring #6). However, not all rings contain the same number of word occurrences (Fig 12(b)): as seen previously in Section 5.1.2, each word occurrence contributes equally and independently to the calculation of the topics distribution. Therefore, a ring containing more word occurrences is more likely to contain more different topics. When we normalise by word occurrences ($\mathcal{N}_{norm}(r)$), the maximum of the normalised topic count (Fig 12(c)) is observed in the first ring. Thus, *ring #1 stands out as the ring that generates proportionally more topics than the other rings*.

In order to validate this hypothesis, we need to rule out that this result is not a mere side effect induced by the structure of the ego networks but it is a tell-tale sign of how humans pick the words in their innermost ring. In other words, we want to test whether keeping the ego network structure unchanged but swapping the words in the rings would still yield the same result regarding ring #1. To this aim, we designed a null model where the ego network structure remains the same but the words are shuffled (more details in the grey box below). In Fig 12(d), we show $\mathcal{N}_{norm}(r)$ for the null model of ego networks. Since the maximum of $\mathcal{N}_{norm}(r)$ is obtained at a different ring r than in the previous case, we can deduce that ring #1 is special not just as a side effect of the ego network structure but due to the nature of the words it contains. To further confirm this finding, note also that the number of topics per word occurrence is significantly lower for innermost rings in the null model with respect to the outermost rings whereas the opposite is true for real ego networks. This is a second element that hints at the peculiar role of innermost rings in real-life ego networks of words.

To extend our study beyond the mere number of topics per ring, we now investigate the diversity in the way topics are distributed, leveraging the entropy of the semantic profiles defined in Section 5.2.1. This is a way of calculating the semantic diversity of the words that compose a ring, as would be a metric like the average pairwise semantic distance, but based on the semantic profile that we have previously calculated. Fig 13 (left) shows different levels of

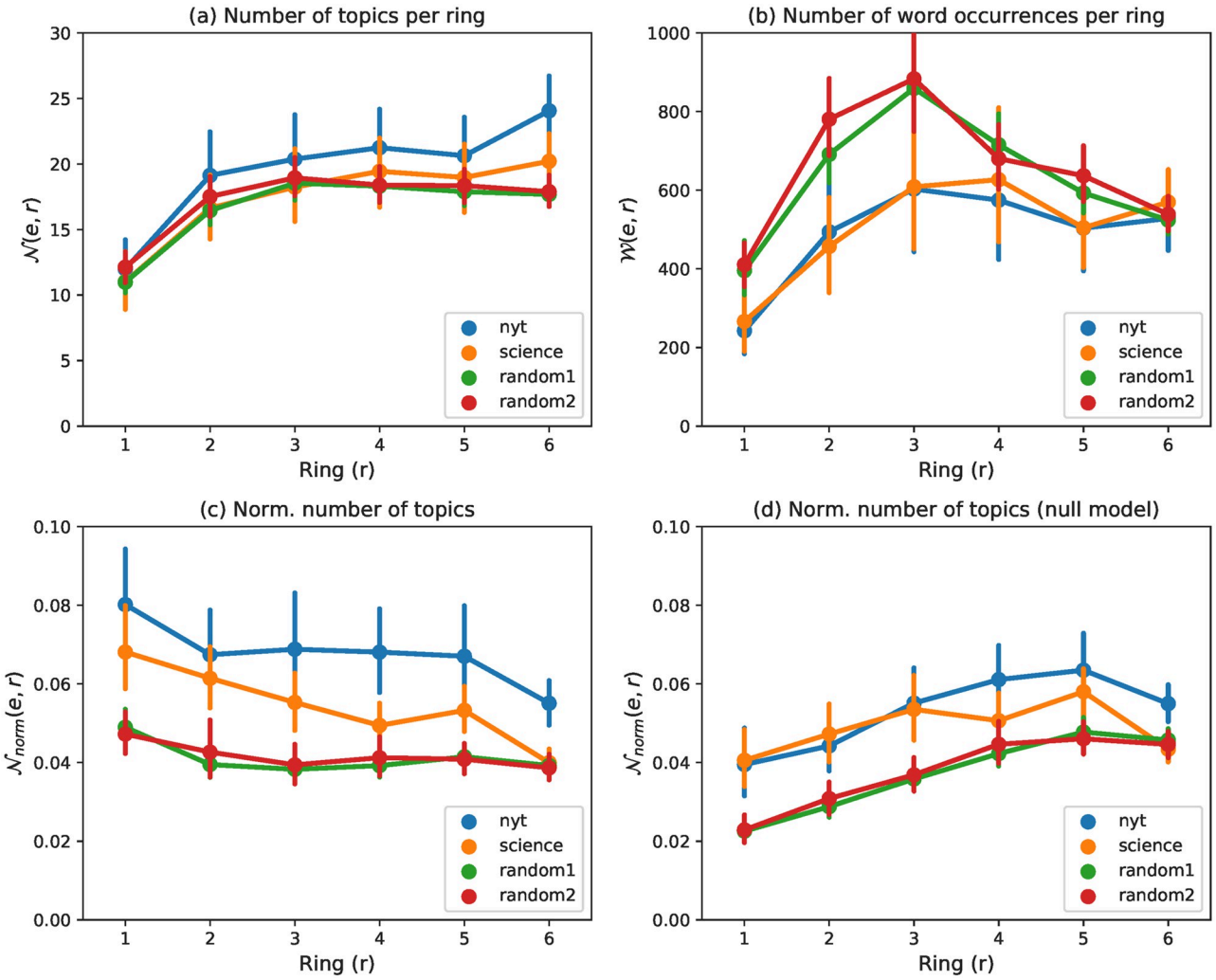

Fig 12. Average number of topics (a), number of word occurrences (b), and normalised number of topics (c) in each ring of the ego network. For "null" ego networks, we report only the normalised number of topics (d).

https://doi.org/10.1371/journal.pone.0277182.g012

entropy depending on the rings: $H(r)$ grows towards the outer rings and is significantly lower in the innermost ring (for all datasets). This means that the outermost rings are, on average, semantically richer than the innermost ones. Then, we compare these results with those obtained from the null model (Fig 13 on the right), to find out whether the differences in entropy are related to the intrinsic structure of the ego network. We find that the entropy of the null model is the same as the original model for all rings, but for ring #1, where the null model entropy is lower. *This means that, even if words are organized in the ego network such that the diversity of topics grows toward the outermost rings, the diversity in ring #1 is higher than what we could expect if words were randomly assigned to rings*, which is consistent with the previous findings of this section.

Building a null model of an ego network.

In order to show that the result is not only determined by the structure of the ego network (independently of the word organization inside), we chose to build "null", artificial ego networks based on those already existing. Let $o(w_u, e)$ be the number of occurrences of the word

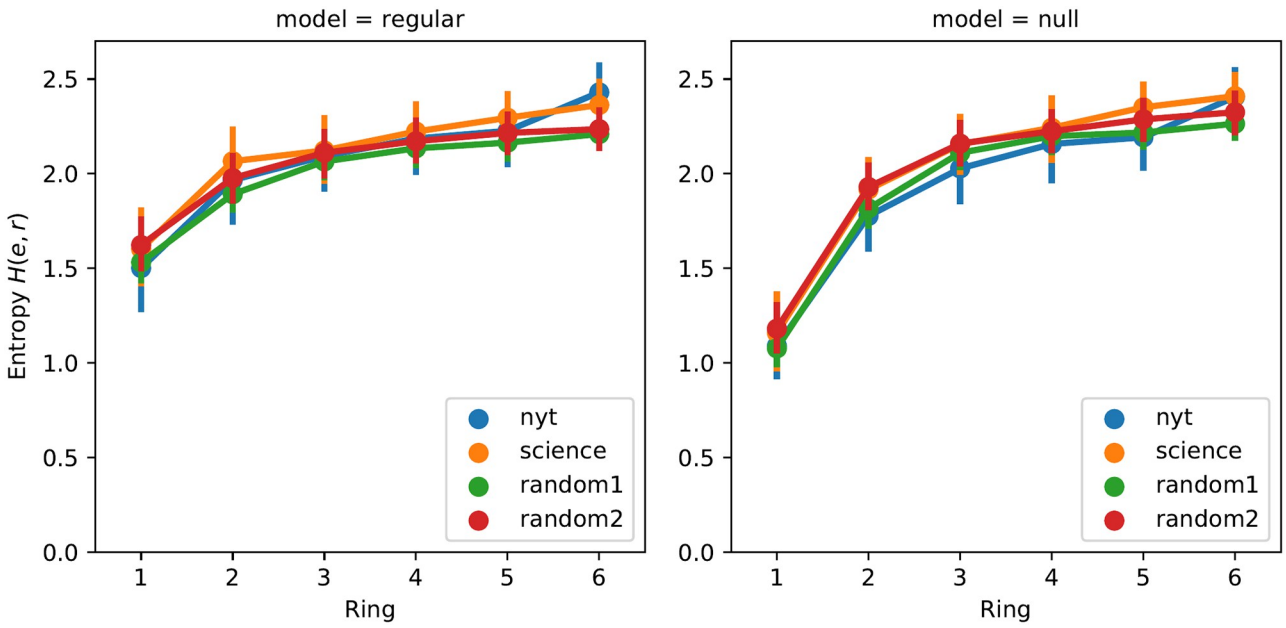

Fig 13. Entropy of the semantic profiles per ring. Real-life ego networks (left) vs null model ego networks (right).

https://doi.org/10.1371/journal.pone.0277182.g013

w_u in ego e, such that the number of word occurrences in a ring r of a given ego e is defined as:

$$O(e,r) = \sum_{w_u \in W_u(e,r)} o(w_u, e), \tag{13}$$

$\mathcal{W}_u(e,r)$ being the set of unique words in ring r. For each ego network, all the words are shuffled (i.e., a new $\mathcal{W}'_u$ is defined) and the word occurrences are artificially changed (new o' and O' are defined) such that the ring sizes and the number of occurrences are kept unchanged:

$$\begin{cases} |\mathcal{W}'_u(e,r)| = |\mathcal{W}_u(e,r)| \\ O'(e,r) = O(e,r). \end{cases} \tag{14}$$

The shuffling process can be considered as a succession of random swaps of words in the ego network. Let us consider a word w_x with X occurrences in ring r_x, and another word w_x with Y occurrences in ring r_y. During the shuffling process, assume the two words are swapped. In that new ego network, the number of occurrences of w_x is forcibly set to the original number of occurrences of w_y and vice versa:

$$\begin{cases} o'(w_x, e) = o(w_y, e) = Y \\ o'(w_y, e) = o(w_x, e) = X. \end{cases} \tag{15}$$

That way, we can preserve Eq (14). Words are shuffled along with their topic distribution $P^{(e)}_{w_u}$ in the original dataset. This topic distribution associated to a unique word w_u is calculated based on its occurrence $w \in \mathcal{W}(e, w_u)$. Each of these word occurrences w is associated with a topic $c_w \in \mathcal{C}$ such that $P_{m(wc)}(c) = 1$. Hence, $P^{(e)}_{w_u}(c)$ simply corresponds to the ratio of the

occurrences of w_u that are associated to c.

$$P^{(e)}_{w_u}(c) = \frac{1}{|\mathcal{W}(e, w_u)|} \sum_{w \in \mathcal{W}(e, w_u)} P_{m(w)}(c). \tag{16}$$

Then the new topic distribution of a given ring r is the weighted average of the topic distribution $P^{(e)}_{w_u}$ of the unique words $w_u \in \mathcal{W}'_u(e, r)$ that compose that ring after shuffling

$$P^{(e)}_{r}(c) = \frac{\sum_{w_u \in \mathcal{W}'_u(e,r)} o'(w_u) \times P^{(e)}_{w_u}(c)}{\sum_{w_u \in \mathcal{W}'_u(e,r)} o'(w_u)}. \tag{17}$$

The full process is summarized with a toy example in Fig 14.

We now carry out a pairwise comparison of the semantic profiles of rings, using the JS distance described in Section 5.2.2. we plot the, in Fig 15. As one can expect, the diagonal is filled with zeros since the distance is calculated between two identical semantic profiles, and the upper triangle mirrors the lower triangle since the distance is symmetric. All datasets exhibit the same features:

- The first row and column always contain the higher values. This means that ring #1 (*i.e.* the innermost ring) is always the most distant from the other rings. In other words, *ring #1 is the most characteristic ring.*
- The lower values are always the distance between ring #5 and #6. Thus, *the pairs of most similar rings are always among the outermost ones.*

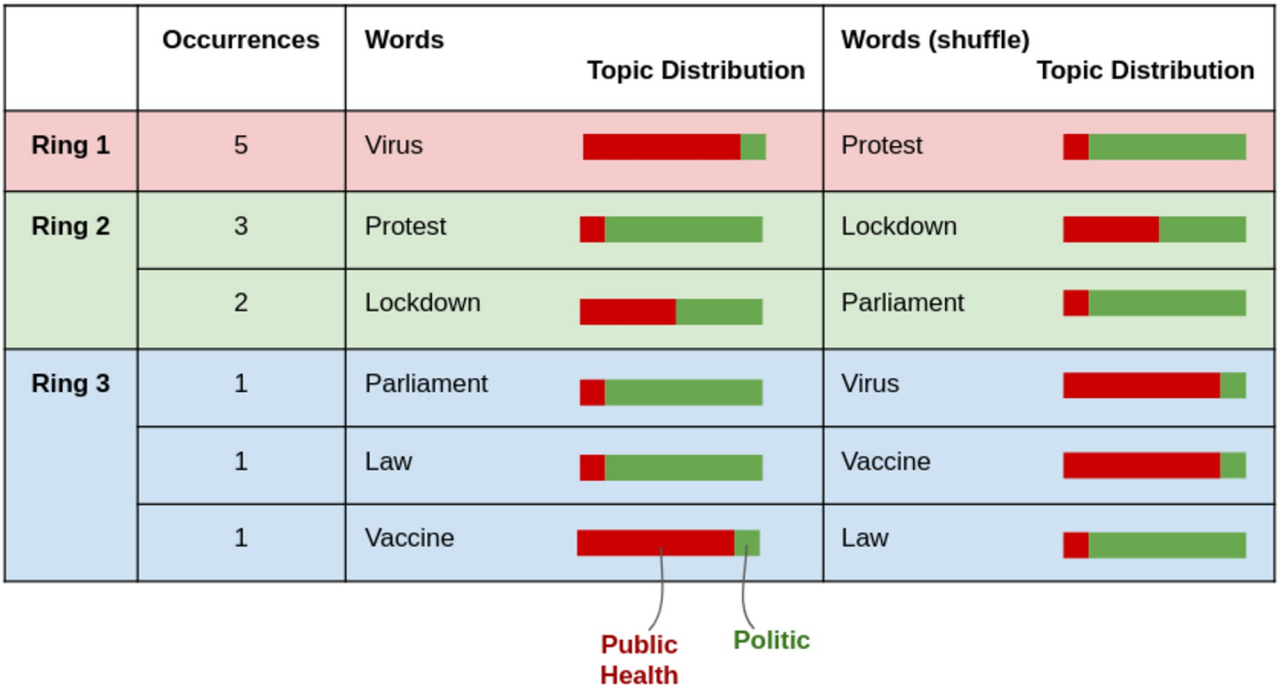

Fig 14. Null model example. The ring sizes and word occurrences are kept, the words are shuffled. In this toy example: $O(e, r_2) = 3 + 2$, $o(virus, e) = 5$, $o'(virus, e) = 1$.

https://doi.org/10.1371/journal.pone.0277182.g014

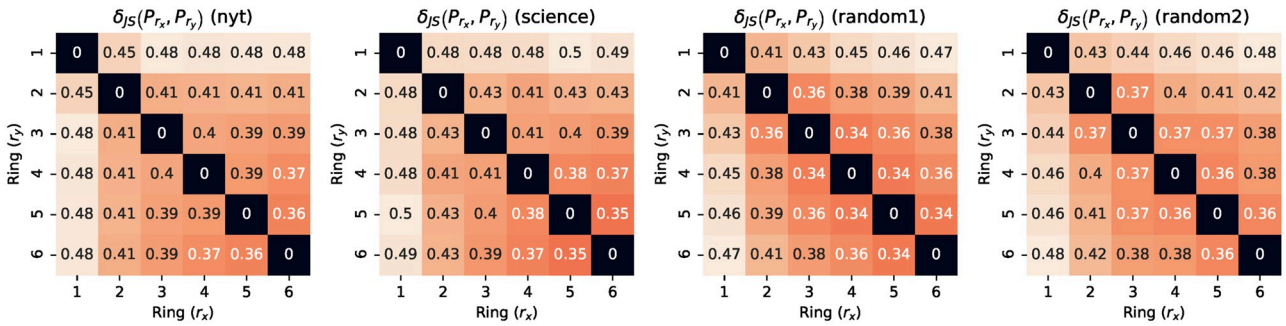

Fig 15. Jensen-Shannon distance. Average JS distance between the rings.

https://doi.org/10.1371/journal.pone.0277182.g015

- For one row or column, the lowest value is always neighbouring the diagonal: given one ring x, the least distant ring is always the previous ring $x-1$ or the following one $x+1$. This means that *two rings close to each other are more likely to be similar*.

The first observation is very important because it shows that the topic distribution associated with the most used words (those in the innermost ring) by a Twitter user is different from that associated with the least used words. This makes ring #1 unique in two ways. *It generates proportionally more topics than the others rings* (Fig 12(c)), *but the distribution in ring #1 is the furthest away from the others* (Fig 15). This hints at a significantly higher "semantic generative role" of inner rings as opposed to outer ones: each word occurring in an inner ring is able "generate" more topics on which the user engages. And these topics, on which that user focuses most (inner rings feature higher frequency of use of words) generate a distribution that is quite distinct from the one at the outermost rings, on which the user engages far less.

Take home message for Section 5.3.1: Ring #1 is special in the ego network of words: it generates proportionally more topics than the other rings, its topic diversity is proportionally higher than expected, and its semantic profile is the most different with respect to the other rings. This suggests that ring #1 may be the *semantic fingerprint* of the ego network of words.

5.3.2 The role of primary topics from ring #1. In the previous section, we discovered that ring #1 is special. It, therefore, makes sense to investigate which topics are most important in this ring and if they tend to be equally important in the other rings. This will allow the reader to familiarize themselves with the methodology as well, before generalizing the analysis to other rings in Section 5.3.3.

We measure the overall importance of r_1's primary topics in another ring r_y by computing $K^{r_y}_{TOP(r_1)}$ (see Section 5.2.3), varying r_y from innermost to outermost layer. Fig 16 shows the coverage of r_1's primary topics in the other rings, across all the ego networks. $K^{r_y}_{TOP(r_1)}$ corresponds to the blue bars in the figure. $K^{r_y}_{TOP(r_1)}$ accounts for approximately 50% of each ring and of the whole ego network (last bar). This small (5–6, on average) set of topics, which fills almost the entire innermost ring, is playing a big role in the entire ego network as well.

To verify if the reverse statement is true (i.e., if topics that are important in the whole ego network are also important in ring #1), we build a new set of topics U_e grouping the most important topics in the whole ego network and calculate $K^{r_y}_{TOP(e)}$. Fig 17 highlights the coverage of those topics across the rings. Although, in general, all primary topics at the level of the ego network are well represented in all rings, we observe a slight predominance in ring #1, as the innermost ring contains the biggest share of the most important topics of the ego network.

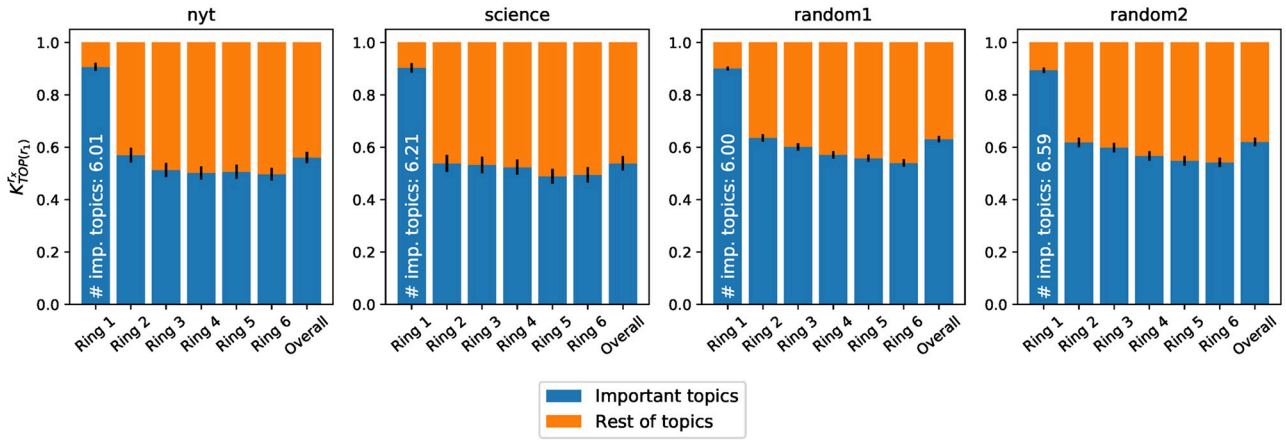

Fig 16. Average strength of ring #1's important topics in the semantic profile of each ring and of the whole ego network. Each bar stands for the semantic profile of each ring (and overall ego network, in the last bar), where the blue part represents the share covered by the most important topics of ring #1 (their average number $|U_{r_1}|$ is written in white).

https://doi.org/10.1371/journal.pone.0277182.g016

This means that topics that are important to the ego network are over-represented in the innermost ring, i.e., an important topic discussed by a Twitter user is very likely to belong to $U_e^{r_1}$.

Take home message for Section 5.3.2: Both results from Figs 16 and 17 indicate a close relation between important topics in ring #1 and those important for the whole ego network. This observation is all the more interesting as ring #1 is semantically the most different from all the others (Section 5.3.1), confirming the special role of this ring in the ego network of words.

5.3.3 Pulling power of primary topics. Let us now focus on the primary topics in a generic ring r_x (i.e., those in $U_{r_x}^{(e)}$). They can also appear in another ring r_y, and can be found in either $U_{r_y}^{(e)}$ or $L_{r_y}^{(e)}$. In the first case, the topics are primary in both rings, in the latter they are primary only in r_x.We now tackle the following problem: which is the ring whose primary topics are most dominant among the primary topics of another ring? This involves measuring the strength, in the semantic profile of r_y, of the topics that are important for both r_y and r_x. Using

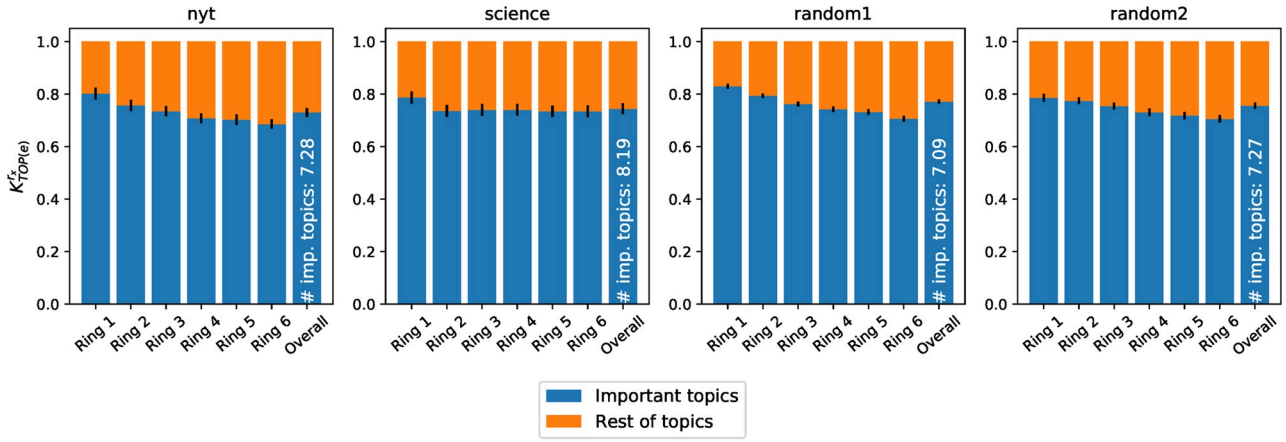

Fig 17. Average strength of the ego network's important topics in the semantic profile of each ring. The blue part of the stacked bar represents the share covered by the important topics in U_e. The average number of topics $|U_e|$ is specified in white.

https://doi.org/10.1371/journal.pone.0277182.g017

Table 6. Pulling power of primary topics. On the left, $S^{r_y}_{TOP(r_x,r_y)}$ for all r_x, r_y pairs in our datasets. On the right, $S^{r_y}_{TOP(r_x),BOTTOM(r_y)}$. In bold, the highest value per column, corresponding to the r_x for which the pulling power is higher in r_y.

	Journalists											
$r_x r_y$	$S^{r_y}_{TOP(r_x,r_y)}$						$S^{r_y}_{TOP(r_x),BOTTOM(r_y)}$					
$\downarrow\rightarrow$	r_1	r_2	r_3	r_4	r_5	r_6	r_1	r_2	r_3	r_4	r_5	r_6
r_1		**.255**	**.226**	**.204**	.195	**.180**		.021	.022	**.023**	.022	**.023**
r_2	.335		.216	.203	.192	.173	.025		**.027**	**.023**	.030	.022
r_3	**.336**	.220		.171	**.196**	.162	**.026**	**.023**		.022	**.032**	.020
r_4	.321	.230	.190		.167	.154	.023	.022	**.027**		.029	.022
r_5	.307	.235	.209	.184		.151	.026	**.023**	**.027**	**.023**		.022
r_6	.318	.234	.210	.188	.179		.025	.024	**.027**	**.023**	.029	
	Science Writers											
r_1		.194	**.191**	**.179**	**.169**	**.158**		.023	**.023**	.027	.027	.023
r_2	.278		.166	.175	.149	.146	**.030**		.022	.025	.027	**.024**
r_3	.285	.172		.154	.153	.146	.026	**.026**		.024	.028	**.024**
r_4	.259	**.200**	.169		.147	.148	.027	.023	.021		.027	**.024**
r_5	**.303**	.180	.183	.168		.141	.027	**.026**	.022	**.028**		.023
r_6	.253	.193	.183	.171	.150		.025	.027	.022	.027	**.029**	
	Random Users #1											
r_1		**.248**	**.216**	**.202**	**.203**	**.190**		**.026**	.024	.026	.026	.026
r_2	**.284**		.202	.192	.189	.178	**.030**		.025	.027	.026	**.028**
r_3	.271	.226		.182	.180	.172	.028	**.026**		**.028**	.026	.027
r_4	.259	.214	.188		.177	.168	.027	.025	**.026**		**.027**	.027
r_5	.267	.211	.193	.181		.168	.028	.025	**.026**	.027		.026
r_6	.260	.213	.189	.175	.171		.028	.023	**.026**	.027	.026	
	Random Users #2											
r_1		**.222**	.199	**.199**	**.179**	**.181**		.024	.021	.025	.020	.025
r_2	**.271**		**.203**	.187	.177	.178	.026		.021	.025	.022	.025
r_3	.250	.213		.184	.169	.178	.025	**.025**		**.026**	.021	.025
r_4	.255	.202	.191		.168	.165	**.027**	.024	**.023**		**.023**	**.026**
r_5	.240	.199	.187	.175		.163	.025	.023	.022	.025		.025
r_6	.246	.207	.190	.178	.158		.023	.023	.021	.024	.022	

https://doi.org/10.1371/journal.pone.0277182.t006

the notation of Section 5.2.3, this is equivalent to studying $S^{r_y}_{TOP(r_x,r_y)}$ for all possible pairs of r_x, r_y. We show $S^{r_y}_{TOP(r_x,r_y)}$ on the left side of Table 6. The diagonal is left blank for the sake of clarity (we are interested in the results when $r_x \neq r_y$). For a given r_y, the largest value is written in bold. We can clearly observe that the primary topics that are also primary in r_1 have almost always the largest share in the semantic profiles of the rings. Beyond the fact that the sum of important topics in ring #1 is also important in the other rings (Section 5.3.2), the table shows that they are on average the most likely to be important in all the other rings.

Now we tackle the complementary question: what is the pulling power of primary topics in a ring on the non-primary topics in another ring? We measure this via $S^{r_y}_{TOP(r_x),BOTTOM(r_y)}$, which is shown in the right part of Table 6.

From the left side of Table 6, we know which is the ring whose primary topics have the highest pulling power on the primary topics of others. But do they have a higher than average strength with respect to the primary topics in the ring as a whole (i.e., regardless of whether they are primary in other rings or not)? To investigate this problem, we show $\sigma^{r_y}_{TOP(r_x,r_y)}$ in

Table 7. Pulling power of primary topics that are also primary elsewhere vs "average" primary / nonprimary topic. On the left, $\sigma^{r_y}_{TOP(r_x,r_y)}$ for all r_x, r_y pairs in our datasets. On the right, $\sigma^{r_y}_{TOP(r_x),BOTTOM(r_y)}$. The highest value per column is in bold.

	Journalists											
$r_x r_y$	$\sigma^{r_y}_{TOP(r_x,r_y)}$						$\sigma^{r_y}_{TOP(r_x),BOTTOM(r_y)}$					
$\downarrow\rightarrow$	r_1	r_2	r_3	r_4	r_5	r_6	r_1	r_2	r_3	r_4	r_5	r_6
r_1		**.059**	**.057**	**.068**	**.051**	**.058**		**.006**	.007	**.006**	.004	**.005**
r_2	.082		.044	**.060**	.043	.051	.006		**.010**	.005	.006	.004
r_3	**.090**	.035		.040	.036	.039	.003	**.006**		.004	**.007**	.003
r_4	.061	.040	.018		.021	.031	.003	**.006**	.009		.006	.004
r_5	.052	.033	.031	.036		.028	.005	**.006**	**.010**	.004		.003
r_6	.061	.032	.027	.029	.018		.004	.005	.008	.005	.004	
	Science Writers											
r_1		.024	**.048**	**.038**	**.043**	**.041**		.002	**.004**	**.006**	.004	**.004**
r_2	**.035**		.033	.027	.022	.025	**.004**		.003	.005	.003	**.004**
r_3	.034	**.025**		.019	.027	.026	.000	.003		.003	.004	.003
r_4	.019	**.025**	.034		.019	.027	.003	.002	.003		.003	**.004**
r_5	.045	.022	.037	.020		.021	.000	.002	.003	.004		.003
r_6	.025	.023	.036	.022	.022		.002	**.004**	**.004**	**.005**	.005	
	Random Users #1											
r_1		**.063**	**.059**	**.049**	**.061**	**.053**		**.006**	.004	.006	.004	.002
r_2	**.061**		.045	.041	.047	.042	**.004**		.005	.006	.004	**.004**
r_3	.045	.039		.032	.037	.036	**.004**	**.006**		**.007**	**.005**	**.004**
r_4	.035	.033	.032		.034	.031	.003	.005	**.006**		.004	**.004**
r_5	.040	.028	.032	.028		.031	.003	.005	**.006**	.005		**.004**
r_6	.035	.032	.033	.023	.028		**.004**	.004	**.006**	.006	.004	
	Random Users #2											
r_1		**.032**	**.043**	**.040**	**.048**	**.041**		.005	.005	.004	.002	.003
r_2	**.057**		.042	.033	**.048**	.038	.002		.005	.004	**.003**	.002
r_3	.041	.024		.029	.037	.037	.002	**.006**		.004	**.003**	.002
r_4	.042	.026	.034		.037	.031	**.004**	.005	**.006**		**.003**	**.004**
r_5	.029	.019	.025	.020		.023	.002	.005	.005	**.005**		.002
r_6	.031	.022	.029	.024	.026		.001	.005	.004	.003	.002	

https://doi.org/10.1371/journal.pone.0277182.t007

Table 7. In the table, all the numbers are positive. This means that, on average, among the most important topics for a ring r_y, if a topic belongs to the important topics of another ring r_x, its strength will be more likely to be higher than the average strength of generic important topics in r_y. A t-test has been performed to assess whether these differences are statistically significant: in all cases, we obtained p-value $< .001$. On the right side of the table we show $\sigma^{r_y}_{TOP(r_x),BOTTOM(r_y)}$, which captures whether topics that are primary elsewhere but not in r_y tend to have a higher share among the least important topics in r_y. In this case, too, the numbers are positive. It also means that, on average, among the least important topics of a given ring r_y, a topic is more likely to have a higher strength if it belongs to the important topics in another ring r_x. Again, the p-values are smaller than .001, confirming that such results are not due to statistical fluctuations.

Take home message for Section 5.3.3: Studying the role of primary topics, we have learned the following.

- Primary topics from ring #1 tend to dominate among the primary topics of other rings. This shows the pulling power of the innermost ring, confirming its special role in the ego network. Vice versa, primary topics from ring #1 do not seem to dominate among non-primary topics of other rings.
- The topics that are primary in some rings tend to be stronger than average among the primary and non-primary topics in the semantic profile of another ring. This effect is especially acute when considering primary topics from ring #1 with respect to generic primary topics in other rings.

5.3.4 Discussion. The study of the semantic profile of the rings of the ego network confirms the relevance of the ego network of words model. This model allowed us to isolate the specific features of the topics associated with the words in the innermost ring. Indeed, the semantic profile in ring #1 is not only the most unique (the most semantically distant from the others), but it is also characterized by both a larger than expected entropy distribution and number of topics generated, when compared with a null model.The most important topics that ring #1 is composed of are not only a set of important topics in the other rings: for every ring, an important topic is more likely to be predominant if it is also important in the innermost ring. Hence, despite the small number of unique words and word occurrences it contains, the innermost ring strongly "predicts" the most important topics in the entire ego network. *In light of these results, we can conclude that the semantic profile of the innermost ring* r_1 *is also the semantic fingerprint of the whole ego network of words*.

As it has been done with social ego networks (using structural properties to study information diffusion [16], or to perform link prediction [53]), we can use the structural and semantic invariants of the ego network of words to investigate some classical data science problems, with a focus on natural language processing. This semantic fingerprint could be used to identify specific Twitter users, or groups of users, with a non-trivial interest distribution for certain topics (e.g. a mix of important topics in the innermost rings and marginal topics in the outermost rings). It could also be used for link prediction with the assumption that users with the same topic of interest in the innermost ego network circles are more likely to follow one another (this is the principle of homophily) or for the purpose of word recommendation in a typing assistance tool. Since we identified some semantic invariants (eg. the role of important topics in ring #1), we could leverage this property to identify outliers deviating from the standard and detect non-human behaviors. Finally, we could use the fact that ring #1 contains the important topics of the entire ego network to spare some time considering only the words in this innermost ring, within the context of topic mining.

6 Conclusion

Inspired by previous work modeling the cognitive constraints that regulate personal social relations, in this paper, we investigate, through a data-driven approach, whether a regular structure can also be found in the way people use words, as a symptom of cognitive constraints in their mental process.Based on a corpus of tweets written by both regular and professional users, we have shown that, similarly to the social case, a concentric layered structure (which we name "ego network of words") very well captures how an individual organizes their cognitive effort in language production and reveals some structural invariants in the way people organise their own vocabulary. Among these invariants, we can list (i) the number of layers (between 5 and 7), (ii) their regular growth from the center of the word ego network outward (the innermost layer is five times smaller than the following one, for all the other layers their size

approximately double moving outward), (iii) the size of external layers (which is pretty stable, with the two penultimate layers accounting respectively for 30% and 60% of the words in the model, regardless of the total number of layers).

Then, going beyond words as units of language, we performed a semantic analysis of the ego network of words. Each ring of each ego network is described by a semantic profile that captures the topics associated with the words in the ring. We have found that ring #1 has a special role in the model. It is semantically the most dissimilar out of the six, and also the one which generates proportionally the largest number of topics. We also showed that the topics that are important in the innermost ring, also have the characteristic of being predominant in each of the other rings, as well as in the entire ego network. In this respect, ring #1 can be seen as the semantic fingerprint of the ego network of words. Finally, we found that the topics that are primary in some rings tend to be stronger than average among the primary and non-primary topics in the semantic profile of the other rings. This shows that, while layer #1 provides a particularly strong signal about prevalence in the ego networks, weaker signals show a more complex structure of influence among topics "resident" in different layers of the ego network of words.

Supporting information

S1 Appendix. Supplementary information on the structural and semantic analysis of word ego networks. In this appendix we provide additional information regarding the data preprocessing, the soft clustering analysis, and we include additional tables to support the findings in the paper.
(PDF)

Author Contributions

Conceptualization: Kilian Ollivier, Chiara Boldrini, Andrea Passarella, Marco Conti.

Data curation: Kilian Ollivier.

Formal analysis: Kilian Ollivier.

Investigation: Kilian Ollivier.

Methodology: Chiara Boldrini, Andrea Passarella, Marco Conti.

Software: Kilian Ollivier.

Visualization: Kilian Ollivier.

Writing – original draft: Kilian Ollivier, Chiara Boldrini, Andrea Passarella.

References

1. Levelt WJ, Roelofs A, Meyer AS. A theory of lexical access in speech production. Behavioral and brain sciences. 1999; 22(1):1–38. https://doi.org/10.1017/S0140525X99001776 PMID: 11301520
2. Broadbent DE. Word-frequency effect and response bias. Psychological review. 1967; 74(1):1. https://doi.org/10.1037/h0024206 PMID: 5341440
3. Qu Q, Zhang Q, Damian MF. Tracking the time course of lexical access in orthographic production: An event-related potential study of word frequency effects in written picture naming. Brain and language. 2016; 159:118–126. https://doi.org/10.1016/j.bandl.2016.06.008 PMID: 27393929
4. Dunbar R. The social brain hypothesis. Evolutionary Anthropology. 1998; 9(10):178–190. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8

5. Dunbar RIM, Sosis R. Optimising human community sizes. Evolution and human behavior: official journal of the Human Behavior and Evolution Society. 2018; 39(1):106–111. https://doi.org/10.1016/j.evolhumbehav.2017.11.001 PMID: 29333060
6. Hill RA, Dunbar RI. Social network size in humans. Human nature. 2003; 14(1):53–72. https://doi.org/10.1007/s12110-003-1016-y PMID: 26189988
7. Zhou WX, Sornette D, Hill Ra, Dunbar RIM. Discrete hierarchical organization of social group sizes. Proceedings Biological sciences / The Royal Society. 2005; 272(1561):439–444. https://doi.org/10.1098/rspb.2004.2970 PMID: 15734699
8. Dunbar RI, Arnaboldi V, Conti M, Passarella A. The structure of online social networks mirrors those in the offline world. Social networks. 2015; 43:39–47. https://doi.org/10.1016/j.socnet.2015.04.005
9. Haerter JO, Jamtveit B, Mathiesen J. Communication dynamics in finite capacity social networks. Physical review letters. 2012; 109(16):168701. https://doi.org/10.1103/PhysRevLett.109.168701 PMID: 23215144
10. Miritello G, Moro E, Lara R, Martínez-López R, Belchamber J, Roberts SGB, et al. Time as a limited resource: Communication strategy in mobile phone networks. Social Networks. 2013; 35(1):89–95. https://doi.org/10.1016/j.socnet.2013.01.003
11. Gonçalves B, Perra N, Vespignani A. Modeling users' activity on twitter networks: Validation of dunbar's number. PloS one. 2011; 6(8):e22656. https://doi.org/10.1371/journal.pone.0022656 PMID: 21826200
12. Sutcliffe A, Dunbar R, Binder J, Arrow H. Relationships and the social brain: integrating psychological and evolutionary perspectives. British journal of psychology. 2012; 103(2):149–168. https://doi.org/10.1111/j.2044-8295.2011.02061.x PMID: 22506741
13. Dunbar R. Theory of mind and the evolution of language. Approaches to the Evolution of Language. 1998;.
14. Brysbaert M, Stevens M, Mandera P, Keuleers E. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. Frontiers in Psychology. 2016; 7(JUL):1116. https://doi.org/10.3389/fpsyg.2016.01116 PMID: 27524974
15. Zipf GK. Human behavior and the principle of least effort. Addison-Wesley Press; 1949.
16. Arnaboldi V, Conti M, Passarella A, Dunbar RI. Online social networks and information diffusion: The role of ego networks. Online Social Networks and Media. 2017; 1:44–55. https://doi.org/10.1016/j.osnem.2017.04.001
17. Ollivier K, Boldrini C, Passarella A, Conti M. Structural Invariants in Individuals Language Use: The "Ego Network" of Words. In: Aref S, Bontcheva K, Braghieri M, Dignum F, Giannotti F, Grisolia F, et al., editors. Social Informatics. Cham: Springer International Publishing; 2020. p. 267–282.
18. Piantadosi ST. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review. 2014; 21(5):1112–1130. https://doi.org/10.3758/s13423-014-0585-6
19. Anderson JR, Schooler LJ. Reflections of the environment in memory. Psychological science. 1991; 2(6):396–408. https://doi.org/10.1111/j.1467-9280.1991.tb00174.x
20. Graesser A, Mandler G. Limited processing capacity constrains the storage of unrelated sets of words and retrieval from natural categories. Journal of Experimental Psychology: Human Learning and Memory. 1978; 4(1):86.
21. Aramaki E, Shikata S, Miyabe M, Kinoshita A. Vocabulary size in speech may be an early indicator of cognitive impairment. PloS one. 2016; 11(5):e0155195. https://doi.org/10.1371/journal.pone.0155195 PMID: 27176919
22. Abel F, Gao Q, Houben GJ, Tao K. Analyzing user modeling on twitter for personalized news recommendations. In: international conference on user modeling, adaptation, and personalization. Springer; 2011. p. 1–12.
23. Bhattacharya P, Zafar MB, Ganguly N, Ghosh S, Gummadi KP. Inferring user interests in the twitter social network. In: Proceedings of the 8th ACM Conference on Recommender systems; 2014. p. 357–360.
24. Frasincar F, Borsje J, Levering L. A semantic web-based approach for building personalized news services. International Journal of E-Business Research (IJEBR). 2009; 5(3):35–53.
25. Arslan O, Xing W, Inan FA, Du H. Understanding topic duration in Twitter learning communities using data mining. Journal of Computer Assisted Learning. 2022; 38(2):513–525. https://doi.org/10.1111/jcal.12633
26. Guille A, Favre C. Mention-anomaly-based event detection and tracking in twitter. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). IEEE; 2014. p. 375–382.

27. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F. Botornot: A system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web; 2016. p. 273–274.

28. Varol O, Davis CA, Menczer F, Flammini A. Feature engineering for social bot detection. In: Feature engineering for machine learning and data analytics. CRC Press; 2018. p. 311–334.

29. Boldrini C, Toprak M, Conti M, Passarella A. Twitter and the press: an ego-centred analysis. In: Companion Proceedings of the The Web Conference x2019;18; 2018. p. 1471–1478.

30. Diaz MT, McCarthy G. A comparison of brain activity evoked by single content and function words: an fMRI investigation of implicit word processing. Brain research. 2009; 1282:38–49. https://doi.org/10.1016/j.brainres.2009.05.043 PMID: 19465009

31. Friederici AD, Opitz B, Von Cramon DY. Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. Cerebral cortex. 2000; 10(7):698–705. https://doi.org/10.1093/cercor/10.7.698 PMID: 10906316

32. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing; 2017.

33. Loper E, Bird S. Nltk: The natural language toolkit. arXiv preprint cs/0205028. 2002;.

34. Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on information theory. 1975; 21(1):32–40. https://doi.org/10.1109/TIT.1975.1055330

35. Jenks GF. Optimal data classification for choropleth maps. Department of Geographiy, University of Kansas Occasional Paper. 1977;.

36. MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1. Oakland, CA, USA; 1967. p. 281–297.

37. Perfetti CA, Wlotko EW, Hart LA. Word learning and individual differences in word learning reflected in event-related potentials. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31(6):1281. PMID: 16393047

38. Senel L K UI, Yucesoy V KA, T C. Semantic Structure and Interpretability of Word Embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018;.

39. Jonnalagedda N, Gauch S. Personalized news recommendation using twitter. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). vol. 3. IEEE; 2013. p. 21–25.

40. Abu-Salih B, Wongthongtham P, Chan KY. Twitter mining for ontology-based domain discovery incorporating machine learning. Journal of Knowledge Management. 2018;. https://doi.org/10.1108/JKM-11-2016-0489

41. Mežnar S, Bevec M, Lavrač N, Škrlj B. Link Analysis meets Ontologies: Are Embeddings the Answer? arXiv preprint arXiv:211111710. 2021;.

42. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;.

43. Grootendorst M. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.; 2020. Available from: https://doi.org/10.5281/zenodo.4381785.

44. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018;.

45. McInnes L, Healy J. Accelerated Hierarchical Density Based Clustering. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). 2017.

46. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 19–27.

47. Radovanovic M, Nanopoulos A, Ivanovic M. Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research. 2010; 11(sept):2487–2531.

48. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9(11).

49. Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory. 1991; 37(1):145–151. https://doi.org/10.1109/18.61115

50. Osterreicher F, Vajda I. A new class of metric divergences on probability spaces and its applicability in statistics. Annals of the Institute of Statistical Mathematics. 2003; 55(3):639–653. https://doi.org/10.1007/BF02517812

51. Jenks GF. The data model concept in statistical mapping. International yearbook of cartography. 1967; 7:186–190.

52. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987; 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

53. Toprak M, Boldrini C, Passarella A, Conti M. Harnessing the Power of Ego Network Layers for Link Prediction in Online Social Networks. IEEE Transactions on Computational Social Systems. 2022;. https://doi.org/10.1109/TCSS.2022.3155946