Extracting *Active* "Ego Networks" of Words: Methodology, Robustness, and Cross-Domain Validation

Kilian Ollivier[1 [0000−0003−2881−5845]] (✉), Chiara Boldrini[1 [0000−0001−5080−8110]], Andrea Passarella[1 [0000−0002−1694−612X]], and Marco Conti[1 [0000−0003−4097−4064]]

CNR-IIT, Via G. Moruzzi 1, 56124, Pisa, Italy
{kilian.ollivier,chiara.boldrini,andrea.passarella,marco.conti}@iit.cnr.it

Abstract. The "ego network of words" model captures structural properties in language production associated with cognitive constraints. While previous research focused on the layer-based structure and its semantic properties, this paper argues that an essential element, the concept of an *active network*, is missing. Drawing inspiration from social ego networks, where the active part includes relationships regularly nurtured by individuals, we establish the notion of an active ego network of words. We demonstrate that without the active network concept, an ego network becomes vulnerable to the amount of data considered, leading to the disappearance of the layered structure in larger datasets. To address this, we define a methodology for extracting the active part of the ego network of words and validate it using interview transcripts and tweets. The robustness of our method to varying input data sizes and temporal stability is demonstrated. In addition, our results are well-aligned with prior analyses of the ego network of words, where the limitation of the data collected led automatically (and implicitly) to approximately consider the active part of the network only. Moreover, the validation on the transcripts dataset (MediaSum) highlights the generalizability of the model across diverse domains and the ingrained cognitive constraints in language usage.

Keywords: ego network of words, active network, cognitive constraints, language production, structural properties

1 Introduction

Human language production is subject to many cognitive processes that unfold transparently. These processes exploit our cognitive abilities (subject to physiological limits such as the duration and volume of long-term memorization of the mental lexicon) to their full extent. For example, it is possible to find the word that best fits the idea that needs to be expressed among thousands of words in only a few milliseconds [16], thanks to complex processing levels (semantic, syntactic, and lexical) involved in speech-related cognition [7]. The structure of the language is influenced by these cognitive strategies. For instance, in most

of the still existing languages, the most frequent words of a language are both the shortests [4] and the most quickly retrieved ones in a speech production task [5,21]. According to Zipf, some of these structural regularities are the result of a compromise that minimizes the effort spent in communication for both the sender – who prefers to use frequent words to minimize the word retrieval time – and the receiver – who prefers less used words to minimize ambiguity. Previous work has shown the existence of a new set of structural [19] and semantic [20] invariants in language production using an egocentric model derived from the social ego network model [3], which in turns originates from the social brain hypothesis from anthropology [10]. This model organizes a person's (the ego) social relationships into concentric circles (between four and five on average) according to their intensity. Recent work has leveraged large amounts of data from social networks to show that this model is also relevant for describing online relationships [12].

In this paper, we adopt a similar approach to study cognitive limitations in language production. Indeed, an ego-centered model organized in concentric layers (called *"ego network of words"*) can be used to describe the way a person uses his personal vocabulary. Language production, just like the socialization process, consumes cognitive capacities that are limited, despite the power of the human brain. These two human activities are closely connected, as postulated by the "social gossip theory of language evolution" [10] which establishes a causal link between the sudden increase in the number of active relationships in humans (from 50 for the closest non-verbal primates to 150 for humans) and the appearance of language that would have optimized the activity of social grooming. Moreover, we expected to find traces of cognitive limits in ego networks of words since we already have evidence of such limits in language production, like the size of the vocabulary which would be about 42,000 words for a 20-year-old native English speaker, or the approximate time span of 180 ms to retrieve a word which is a strong constraint [8].

1.1 Contributions and key results of the paper

The ego network of words is a novel model that captures structural properties in language production linked to cognitive constraints. Existing works focused on the layer-based structure and its semantic properties. Here, we argue that the model is still missing a key element used in the characterization of social ego network, i.e., the concept of active network. In social ego networks, the active part of the ego network only included relationships that the ego spent time nurturing, thus consuming cognitive resources on the ego's side. The layered structure of the social ego network only emerged in the active part. Such "meaningful" relationships were identified with a shoe-leather anthropology approach, based on a common understanding of how human social interactions work. Specifically, a relationship was considered meaningful if it entailed at least one interaction per

year, based on the fact that people close to each other exchange at least birthday or holiday wishes[1].

In previous works on language ego networks, the layered structure seemed to emerge without applying any preliminary filter in the spirit of the birthday/holiday wishes. And anyway, finding such a common sense threshold for the ego network of words would not have been possible. In this paper, we argue that without the notion of "active" ego network of words, the analysis carried out would not be robust to the amount of data considered. Specifically, in the paper we show three key properties in this regard. First, that depending on the size and extent of collected data, ego network may or may not include (a part of) the inactive ego network. Second, that appropriate filtering is needed, in order to isolate the active part of the ego network. Third, that layered structures – the fingerprint of the human cognitive involvement – emerge only when the inactive part of the ego network is excluded. Therefore, the paper provides evidence about the complete structure of the ego network of words, as well as a robust methodology to isolate and study it.

The first contribution of this paper is the definition of a methodology to extract the "active" part of the ego network (Section 4). In Section 5.1, we successfully test this methodology using two types of datasets: interview transcripts and tweets. MediaSum is a dataset that includes thousands of verbatim transcripts of spoken interviews from an American public radio and private TV channel (Section 3.1). The Twitter datasets are extracted from the same users as in [19], but we downloaded larger timelines, up to 10K tweets (Section 3.2). We also prove that the method that we use to extract the active ego network is robust to different amounts of input data (Section 5.3) and that the active size is stable over time (Section 5.4). The structural results (Section 5.2) of the ego networks produced in this way substantially confirm the layer ego network of word structure obtained in previous work [19] but are robust to the size of the input data. The second contribution of the paper is the validation of the ego network of words model on a dataset (MediaSum) that is completely different in nature from the Twitter ones on which it had been applied previously (Section 5.2). The fact that the structural properties of the word ego networks are confirmed is an important validation that the model generalizes across different domains and, thus, that the underlying cognitive constraints are ingrained in our use of language.

The key findings of the article are as follows:

- We introduce the notion of *active part of an ego network of words*, beyond which the model would contain words that are not used frequently enough to denote a cognitive involvement. We show that, beyond the active part, the word ego network becomes poorly structured (*i.e.* with a very low number of concentric circles).
- We define a robust algorithm to extract this active part based on the properties of the ego's language production.

[1] These considerations hold for Western societies, which were the focus of this anthropological studies.

- We find that the active size is specific to each ego network and stable over time. Therefore, each ego appears to have its own limit to the number of words it can actively use, similarly to what was observed for social ego networks.
- Even if the ego networks are larger than those observed in previous papers [19,20] (where the concept of active network was not exploited) we retrieve most of the structural invariants previously observed: first, the number of circles in the model is approximately the same. Second, third-to-last and second-to-last circles account for 30% and 60% of the words in the ego network whatever the number of layers. Third, the scaling ratio between circles tends towards 2.
- Ego networks based on oral language production (interviews) have the same structural properties as those obtained from Tweets, thus confirming the cross-domain generalizability of the ego network model.

2 Related work

2.1 Social ego networks

The social ego network model organizes the interpersonal relationships of a person (the ego) into concentric circles. This is an empirical model derived from the work of anthropologist Robin Dunbar on the number of active relationships that a human can maintain on average over time [10]. To do this, he established a correlation between the relative size of a part of the brain dedicated to sociability (the neocortex) and the typical group size in primates, then deduced what the equivalent number would be for humans. This number, 150, is called Dunbar's number. Anthropological studies have shown that this number is a recurrent occurrence in human organizations, as can be observed in Hutterite communities where it is the maximum number before the group splits up, in Israeli Kibbutzim where it is the average number at the foundation time, but also in modern factories sizes [11]. By analyzing the traces left by online social interactions, researchers have shown that the number of active online relationships that can be maintained at the same time is in the same order of magnitude as the Dunbar's number [12]. Moreover, for a given person (the ego) it is possible to subdivide these active relationships (alters) into four concentric circles [14,23], the most central one containing the most intimate relationships. These circles contain about 5, 15, 50, 150 alters, and exhibit a consistent scaling ratio of three in their sizes. This model of concentric circles, called "ego network model" was also confirmed for online relationships, with approximately the same number of circles and the same scaling ratio [12] as for offline relations. Thanks to online social networks, we also know that after an initial moment of growth, the ego network structure remains stable over time for the majority of the individuals [2,1].

2.2 Structural and semantic properties of ego network of words

Previous papers have shown the relevance of using an ego network of words for studying language production [19,20]. Using datasets extracted from Twitter,

ego networks of words were constructed with a methodology similar to that used to construct social ego networks. However, instead of considering other people as alters and the frequency of contact with the ego as a proxy for the intensity of the relationship, words were considered as alters, and their proximity to the ego is measured by their frequency of use. In this way, each ego's vocabulary is organized into concentric layers, the first of which would contain the most frequently used words while the last would contain the least used words. Even if, unlike social ego networks, the size of the ego network varies significantly, the number of layers remains in the same order of magnitude: between five and seven [19]. A very strong similarity in the relative size of concentric layers between egos with the same amount of layers was found, regardless of the dataset. Moreover, the third-to-last and second-to-last layers account for 30% and 60% of the words in the ego network whatever the number of layers, which means that the total number of layers depends on the number of internal layers (from the innermost to the third to last), which is determined by the distribution of the most frequent words. Finally, it appeared that the scaling ratio is not three as in the case of the social ego network, but tends towards two consistently when moving towards the outer layers. A semantic analysis of the rings was also performed, assigning each one a semantic identity card [20]. This is a distribution of the importance given to one hundred topics found automatically and common to a whole dataset. We found that the innermost ring is the most different from the others, as it generates proportionally more topics. All the important topics of this ring are also important in the whole ego network and vice versa. That is why this layer can be seen as the semantic fingerprint of the ego network.

3 The datasets

In this study, we will rely on two types of datasets. The first, MediaSum [24], compiles years of television and radio interview transcripts. In the second, we collected up to ten thousand tweets each from four distinct groups of Twitter users.

3.1 MediaSum

MediaSum contains about 464K interview transcripts, of which 49K are from NPR (American public radio) and 415K from CNN (cable news channel). These interviews are extracted from well-known broadcasts, such as "Anderson Cooper 360 degrees" on CNN or "Morning Edition" on NPR. This is a valuable dataset, as it allows us to study the ego networks of words produced from spoken-language corpora collected over a long period of time. Indeed, the dataset contains between 10K and 35K interviews per year between 2000 and 2020 (Figure 1). The speakers are mainly television or radio anchors and recurring guests. Another advantage is that the topics of the interviews are diverse (*eg.* politics, international news, crime), and so are the guests such as the athlete Michael Phelps or the actor Morgan Freeman. Each interview lasts on average 30 turns (each turn corresponds

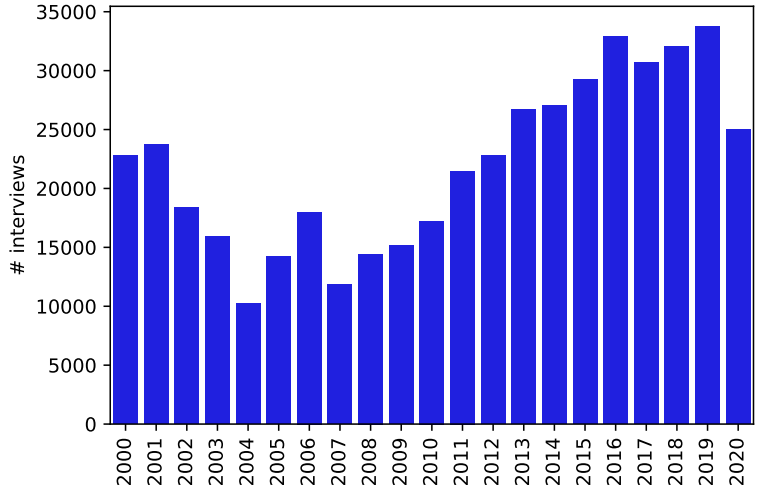


Fig. 1: Number of interview transcripts per year in the MediaSum dataset.

to a speaker's line of dialogue that we call "utterance") and involves 6.5 speakers (4.0 for NPR and 6.8 for CNN). Taking into account its characteristics, this dataset is particularly interesting for investigating the long-term cognitive limitations related to the language of various kinds of people.

Cleaning the dataset

Since we want to group all of the dialogue lines for each person across the entire dataset, we must first clean the names which are manually filled (*eg.* "wozniak", "steve wozniak", "steve wozniak, founder, apple computer", "mr. steve wozniak (co-founder, apple computer)"). After this name-cleaning operation and a first round of deletion of speakers with too few utterances (mainly due to inconsistencies in their names like spelling mistakes), we end up with 106,627 speakers. The average number of utterances per speaker is around 124 (Table 1). In our previous papers [19,20], where we used corpora extracted from Twitter, we defined a minimum of 500 tweets per user. In a similar way, we keep only speakers with at least 500 utterances such that the corpora to process have a minimum size. This criterion results in the suppression of 98.6% of the speakers, but only 55% of the total number of utterances in the dataset. This relatively small group of speakers produces almost half of the text corpus, that we will use to build ego networks of words. The sentences are tokenized, the stop words are removed and the remaining tokens are lemmatized to group together inflected versions of the same word. Once we obtain the number of words' occurrences for a given speaker, we remove those that appear only once to leave out most misspelled words. As we can see in Figure 2 and Figure 3, a few speakers have a very large number of word occurrences and unique words. Unsurprisingly, most

	Before	After
Number of speakers	106,627	1,513
Number of utterances	13,228,854	5,931,363
Number of utterances / speaker	124	3,920
Number of words / speaker	–	89,313
Number of unique words / speaker	–	5,316

Table 1: MediaSum statistics, before and after removing speakers with less than 500 utterances (word stats are only computed for users with $>$ 500 utterances)

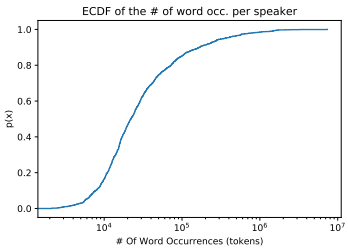


Fig. 2: Word occurrences per speaker

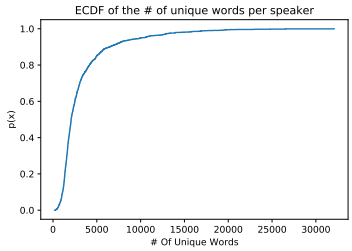


Fig. 3: Unique words per speaker

of them are anchormen or anchorwomen, like Wolf Blitzer of CNN, who are the most active speakers in the dataset. The majority of speakers have between 10K and 100K word occurrences and less than 5K unique words. The average number of word occurrences among all the speakers is 89,313 and the average number of unique words is 5,316.

3.2 Twitter

In [19,20], we built ego networks of words based on Twitter timelines with up to 3.2K tweets (the download limitation of the standard Twitter API) collected from four sets of users:
working for the New York Times. The NYT itself has created a list of 678 accounts [2].
who tweet about science-related topics. A list of 497 accounts has been created by Jennifer Frazer[3], a science writer at *Scientific American*.
are sampled among accounts that published on January 16, 2020 (download time) a tweet or a retweet in English containing the hashtag *#MondayMotivation*. This hashtag, which is both popular and neutral, does not refer to a political or controversial issue. Bot accounts are filtered using the Botometer service [9] which leverages both structural properties (number of followers, tweeting frequency,

[2] `https://twitter.com/i/lists/54340435`
[3] `https://twitter.com/i/lists/52528869`

Dataset	# of users	Avg. word occ. / users	Avg. words/users
NYT journalists	285	87,698	11,877
Science Writers	256	138,050	14,952
Random users #1	1,536	48,021	6,650
Random users #2	1,324	57,177	6,757

Table 2: Twitter datasets after removing users with less than 500 tweets

etc) and language features to detect non-human behaviors. After this operation, the dataset contains 5,183 accounts.
are sampled among accounts that issued on February 11$^{\text{th}}$ 2020 (download date) a tweet or a retweet in English, from the United Kingdom. The group contains 2,733 accounts after removing the bots.
We extended the timelines of these four sets of users to up to 10K tweets, by leveraging the extended download capabilities of the Twitter Academic Research track. As illustrated by Figure 4, this results in much longer timelines with respect to those analysed in previous works. These longer timelines are used to stress-test the ego network of words model. In the same fashion as in [19,20], and in Section 3.1, we only keep the timelines with at least 500 tweets. The figures related to the number of word occurrences and unique words are reported in Table 2. Even if the numbers are lower for both random user datasets compared to journalists and science writers, all figures are of the same order of magnitude as for MediaSum.

4 Methodology

4.1 Preliminaries

Before describing our method for building the ego network of words and extracting its active part, we introduce here the notation used in the section (also summarised in Table 3). We denote an ego with the letter e, where the ego is the speaker (MediaSum) or user (Twitter) in our datasets for whom we want to extract the ego network of words. After the cleaning process discussed in Section 3, for each ego e we end up with a tuple (i.e., an ordered sequence) of tokens [18], which we denote with $\mathcal{T}_e$. Note that the tokens in $\mathcal{T}_e$ are generally not unique. In computational linguistics, the term *type* denotes the class of all tokens containing the same character sequence [18]. In other words, the set of types corresponds to the set of distinct tokens or, slightly simplifying, a type is a word and its occurrences are tokens. For example, in the sentence *a rose is a rose is a rose*, there are eight tokens but only three types. In this paper, for the sake of simplicity, we use the terms *type* and *word* interchangeably. Similarly, tokens may be also called *occurrences*. In the following, we denote the tuple of unique words in an ego

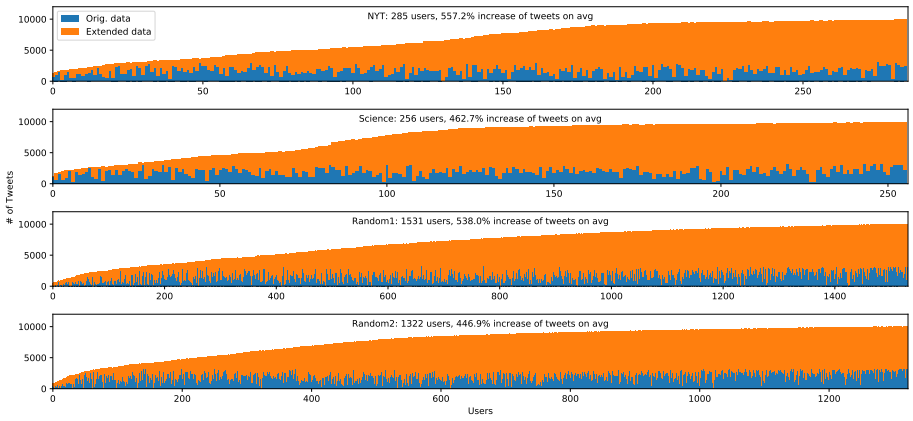


Fig. 4: Collected Twitter timelines containing at least 500 tweets. Each bar corresponds to a timeline, where the blue part refers to the number of tweets in the original dataset, and the orange part refers to the number of newly collected tweets.

network as $\mathcal{W}_e$. Please note that both $\mathcal{T}_e$ and $\mathcal{W}_e$ are ordered sequences, where the order is defined by the appearance in the ego's timeline in chronological order. So, if we observe the first n tokens in the ego's timeline, we will get exactly n tokens but at most n unique words. We denote with $\mathcal{T}_e^n$ and $\mathcal{W}_e^n$ the tuples of tokens and unique words, respectively, observed up to n. We call n_f the maximum value of n (corresponding to the overall number of tokens in the observed timeline for ego e) such that $\mathcal{T}_e^{n_f} = \mathcal{T}_e$ and $\mathcal{W}_e^{n_f} = \mathcal{W}_e$, where $|\mathcal{T}_e| = n_f$ and $|\mathcal{W}_e| \leq n_f$.

In the rest of the section, when there is no risk of ambiguity, we will drop the subscript e from our notation: in that case, all the variables discussed will be referring to the same tagged ego e.

4.2 Legacy method for building an ego network of words

Ego networks of words are used to hierarchise the words used by a given person based on their frequency. In the following, we summarise the model presented in [20]. Let us focus on a tagged ego e (hence, hereafter we drop the subscript e in the notation). The ego network of words model is such that each word from $\mathcal{W}$ is assigned to one of τ rings $r_1, r_2, \ldots, r_\tau$, knowing that r_1 (the innermost ring) contains the most frequently used words and that r_τ (the outermost ring) contains the least used words. The set of words assigned to the ring r_i is called $\mathcal{W}_{r_i}$ such that:

$$\mathcal{W} = \bigcup_{i=1}^{\tau} \mathcal{W}_{r_i}. \tag{1}$$

Symbol	Description
$\mathcal{T}_e$	tuple of tokens, i.e., sequence of words ego e has used
$\mathcal{W}_e$	tuple of unique words used by ego e
$\mathcal{T}_e^n$	$\mathcal{T}_e$ cut at the n-th token
n	length of the tuple $\mathcal{T}_e^n$
$\mathcal{W}_e^n$	unique words in $\mathcal{T}_e^n$
w_e^n	length of the tuple $\mathcal{W}_e^n$
n_f	overall number of tokens in the observed timeline for ego e
n_a	active network cut-off
τ_e	optimal number of circles
r_i	i-th ring of the ego network
l_i	i-th layer of the ego network
$\mathcal{W}_{e,r_i}$	unique words assigned to ring r_i

Table 3: Summary of notation used in the paper.

The ego network can also be studied from a cumulative perspective with concentric layers l_1, l_2, ..., l_τ, with layer l_i containing all the rings r_j where $j \leq i$. The set of words assigned to layer l_i is denoted with $\mathcal{W}_{l_i}$, so:

$$\mathcal{W}_{l_i} = \bigcup_{j=1}^{i} \mathcal{W}_{r_j}. \tag{2}$$

This implies that the innermost layer l_1 is equivalent to r_1.

Words in an ego network are characterized by their usage frequency, which corresponds to their number of occurrences divided by the observation window (which is the same for all words uttered by the same ego). To find the best natural grouping of words (i.e., to find τ) we use the Mean Shift [13] algorithm, which is able, in contrast to Jenks [15] or K-Means [17], to automatically optimize τ, the number of groups to be found. Clustering on a unidimensional variable is equivalent to dividing the word frequencies into mutually exclusive intervals. The Mean Shift algorithm detects clusters that correspond to the local maxima of an estimated density function of word frequencies. This function is obtained with the kernel density estimation for which the sensibility is set with a fixed parameter called the bandwidth. We apply a preliminary log-transformation to frequencies in order to compress high values and ensure that the same bandwidth setting allows peak detection for both the high- and low-frequency parts of the distribution[4]. The obtained clusters of words correspond to the τ rings of the newly built ego network of words for ego e, r_1 being the cluster containing the most frequent words and r_τ the one containing the least frequent words.

[4] Note that applying a log-transformation to word frequencies is common in psychological research when studying the associated cognitive processes [6].

4.3 Motivating the need for an active ego network extraction method

We start by applying the methodology described above to all the words in $\mathcal{W}$ for the egos in our datasets, and we plot the distribution of the number of circles τ in Figure 5. We can observe that the obtained ego networks of words have a very low number of circles (the most frequent case is two) compared with the ego networks of words in previous work (usually between five and seven circles [19,20]), despite exactly the same workflow being used. Note also that the Twitter datasets used here are *the same* as those in [19,20] except for the timeline length considered (much larger, in this work). As we can observe in Figure 6, ego networks with one or two circles are the biggest ego networks (*i.e.* with the largest number of unique words $|\mathcal{W}|$). This seems to suggest that, when considering larger textual inputs, the ego network model loses its finer discriminative power. In fact, two-circle ego networks are considered uninteresting, as they simply separate the most used words from the least used words.

However, this finding is not unexpected: in the social ego network case, the theory distinguishes between the *full* and *active* ego network, stating that only the relationships in the active part are actually consuming cognitive resources [3]. The conventional cut-off point, as stated in [10], is for the social relationship to involve interactions at least once a year, which, in Western societies corresponds to at least exchanging Christmas/birthday wishes. While this cut-off point could be obtained with anthropological common sense for social ego networks, it is difficult to come up with a similar rule of thumb for the ego networks of words, which are less rooted in everyday experiences. Hence, in this work, we set out to design a methodology to automatically extract the cut-off point in the ego networks of words. This methodology should then be applied before building the ego networks as described in Section 4.2, in order to discard the words that do not take up cognitive capacity.

4.4 Extracting the active ego network

The idea behind an *active* ego network is that all the words it contains should be actively used, even those in the outermost circle. If we let a person speak, we notice that from a certain point on the frequency of appearance of a new word decreases rapidly: a specific number of words is sufficient for this person to express him/herself. This quantity is the maximum number of actively used words. We can observe this phenomenon in Figure 7, where the number of tokens $n = |\mathcal{T}^n|$ vs the corresponding number of unique words $w^n = |\mathcal{W}^n|$ is plotted for a single speaker in the Mediasum dataset (we define w^n to improve the readability of the formulas in the following sections). The curve is obtained by scanning the timeline (or, more exactly, the chronologically ordered tokens remaining after preprocessing the timeline) from start to end, and counting the new tokens and the unique words as we go. The catch is that not every new token corresponds to a new unique word. We will call this curve the *saturation curve*, which we denote with s. Using the notation in Section 4.1, $s : n \mapsto w^n$.

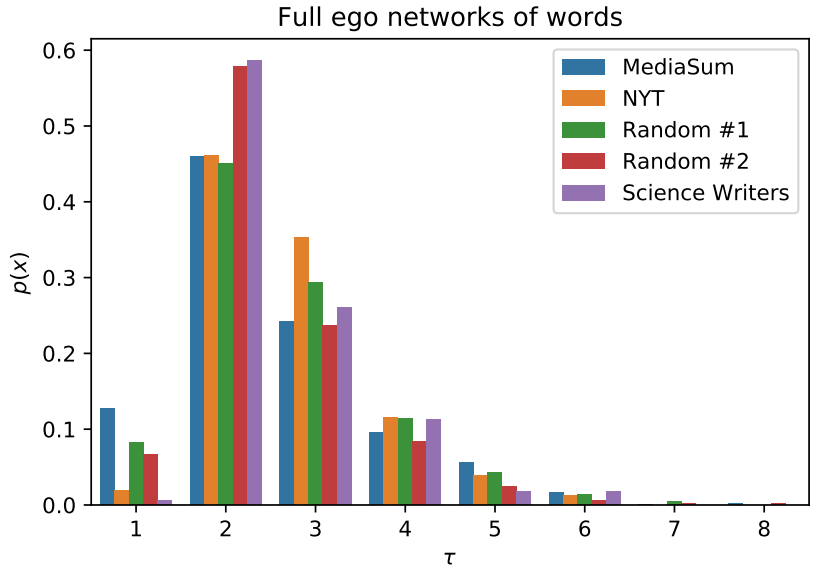


Fig. 5: Distribution of the number of circles τ when considering all the words available in $\mathcal{W}$

In Figure 7 and 8, we present two typical cases observed in our datasets. Figure 7 serves as a representative example of a broad trend that emerges in our data for users who have been observed over an extended period. Initially, there is a swift growth in the number of discovered words as new tokens are explored, but in the second phase, this growth rate significantly decreases. The rate at which new words are discovered remains fairly constant in both phases. Figure 8 is representative of users who were not observed for a sufficient duration to reach the second phase described in Figure 7. In this example, the total number of tokens is much lower, comparable to the number of tokens in the initial phase for users represented in Figure 7. We argue that the active part of the ego network ends at the cut-off point of the saturation curve, i.e., where the first regime ends and the second one begins. The saturation curve shows how many tokens are needed to observe a certain number of unique words. The number of tokens needed to increase the number of words by one can thus be seen as the maximum number of tokens an ego can use without including a new word in his spoken or written expressions. Saturation curves of "mature" ego networks show two regimes, whereby *in the first one* words appear "sooner", meaning that the user is able to "resist" less before "injecting" a new word. Before proceeding further, it is important to acknowledge that in general, non-linear saturation curves may exhibit less regularity than the one depicted in Figure 7, while the overarching pattern of two distinct major regimes remains consistent. This might present a challenge for algorithms intended to automatically identify the transition point between regimes. This is the rationale behind our proposal,

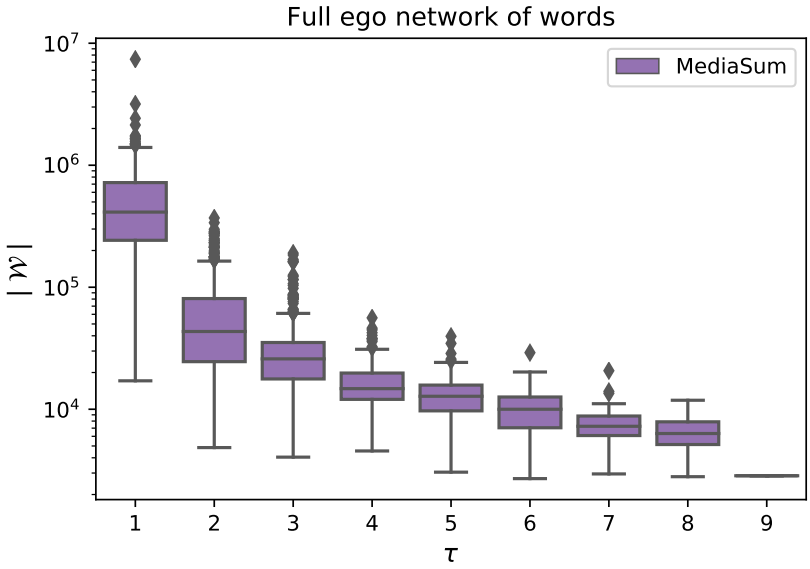


Fig. 6: Number of circles τ vs full ego network size ($|\mathcal{W}|$) in the MediaSum dataset. The same trend is observed in the other datasets (plots omitted to optimize the space).

outlined in Section 4.4, for a recursive algorithm that only terminates when the major trends are identified.

Recalling that the saturation curve is defined as $s : n \mapsto w^n$, the goal of this section is to describe a methodology for finding the value of n (which we call $\hat{n}_a$) where the first phase described above ends and the second one begins. The number of unique words at the cut-off point n_a of the curve corresponds to $w^{n_a} = |\mathcal{W}^{n_a}|$, while $w^{n_f} = |\mathcal{W}^{n_f}|$ corresponds to the total number of unique words in the *full* ego network (n_f being the maximum value of n). If our intuition is confirmed, the well-known layered ego network structure would emerge by considering only words in the first regime of the saturation curve when computing the ego network. Indeed, we show this in Section 5.1. Note that sometimes the textual data for one ego is not large enough for the ego network to reach any cut-off point (Figure 8). This means that the cognitive capacity for language production is not fully exploited (in the textual information available in our datasets), so the ego network of words is not fully formed. In this case, we remove the egos from the analysis because only mature ego networks are reliable for extracting structural properties.

Methodology for identifying the cut-off point. We start with a high-level description of our methodology, illustrated in Figure 9. Let us focus on the curve s, and assume that it is not linear in $[0, |\mathcal{T}|]$ (if it is linear, we can stop searching for the cut-off, since there is none). Our cut-off point n_a would split s

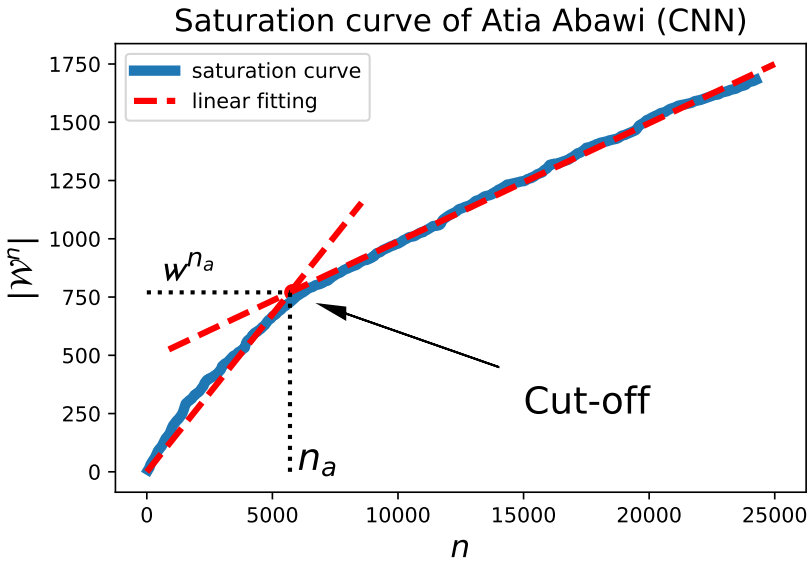


Fig. 7: Non-linear saturation curve.

in two halves: in the first one, s is approximately linear and with a greater slope; after n_a the saturation curve enters a regime of reduced growth (in this second regime, s might be linear or not). We want to find the knee point in s where the slope change is observed. The search for n_a is done recursively, continuing to split the first half until it is effectively linear. At this point, the algorithm stops. The intuition is that the words and tokens before n_a correspond to the first regime described above, where new words are discovered at a higher rate. This recursive approach allows us to discard minor irregularities in the saturation curve and to properly detect the major trend of linear growth.

Algorithm 1 summarises our approach. The recursive search is carried out through the RecursiveCutOff function, which is initially fed all data points from the saturation curve. If the saturation curve is already linear, then the algorithm returns n_f, the upper bound of n. If the saturation curve is not already linear, we need to split it into two halves. We do this with the SplitSaturationCurve function, which tests all the possible cut-off points and selects the one guaranteeing the best (in terms of residual sum of squares) linear fit on both sides of the cut-off. Then, we focus on the linearity of the first half to ensure there is no more potential cut-off (we are not directly concerned with the linearity of the second part, because, as long as we are able to detect a phase change, the second part will be dropped anyway being it outside of the active network). What we want to assess is whether the "signal" in the first part of the saturation curve (before the current cut-off) is *mostly* linear. To this aim, we leverage Lasso regression [22] for its ability to operate a variable reduction on its input features. The features used by Lasso are the polynomial terms of

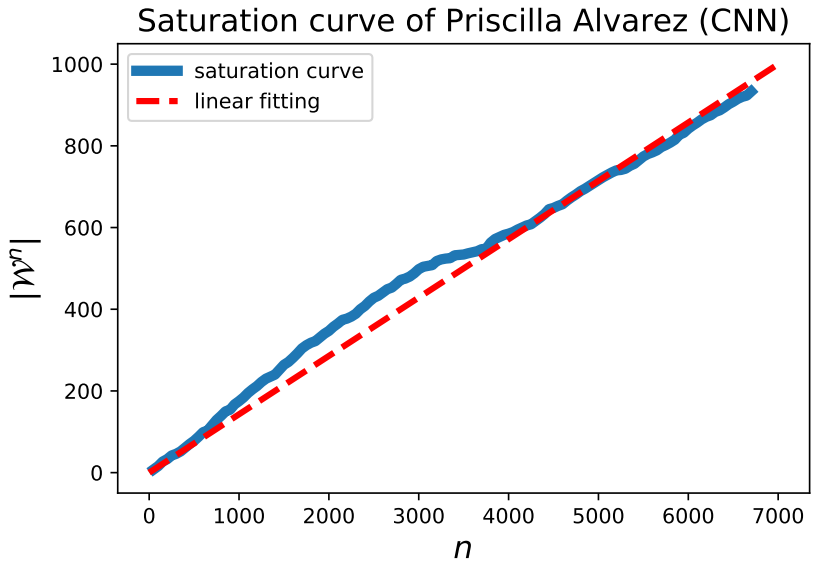


Fig. 8: Linear saturation curve.

the inverse saturation curve (we consider the inverse for ease of explanation). Specifically, we consider the following: $s^{-1}(t) \sim \sum_{i=1,\dots,p} \beta_i w^i$, with β_i being the coefficient optimized by Lasso and $s^{-1}(t)$ the inverse of the saturation curve. In other words, we consider the growth of the number of unique words with respect to the number of tokens, and evaluate whether the dependency is mostly linear, mostly quadratic, etc. Intuitively, in the first regime of the saturation curve, the growth is linear because each new token roughly corresponds to a new unique word. Vice versa, in the second regime, we observe an inflection. Then, with the LassoMaxVariableReduction function, we denote a Lasso regression where the λ parameter for regularization is chosen such that only one coefficient of the regression is set to a non-zero value: the one corresponding to the most significant polynomial term. If the nonzero coefficient corresponds to the linear term, we confirm that the saturation curve before the current cut-off point is linear enough for our purposes, and we stop the search. Once we obtain n_a, we can use it to obtain the active ego network. Specifically, the words in the active ego network of e are $\mathcal{W}_e^{n_a}$.

To summarize, the algorithm returns a value called $\hat{n}$ that corresponds to n_a if there is a cut-off point, and to n_f if there is not. With this algorithm, we can separate the egos into two groups: those that have a mature ego network (i.e., those for which we have been able to extract a cut-off in the saturation curve) and those that do not. The number of egos in the first and second groups is shown in Figure 10 for our datasets. It appears that in all datasets, and especially in the largest ones (MediaSum and both random datasets), egos with mature ego networks are the vast majority. In the rest of our analysis, we will retain only them, so that we can study their structural properties.

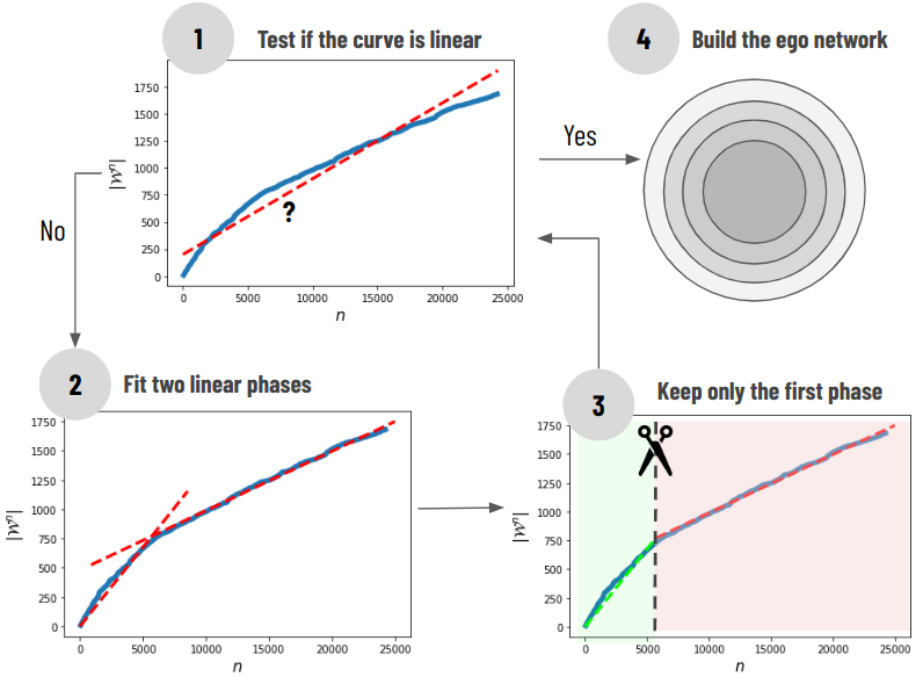


Fig. 9: Steps for detecting the saturation point. 1) Linearity test. 2) If the curve is not linear, we find the best model fit with two linear parts. 3)

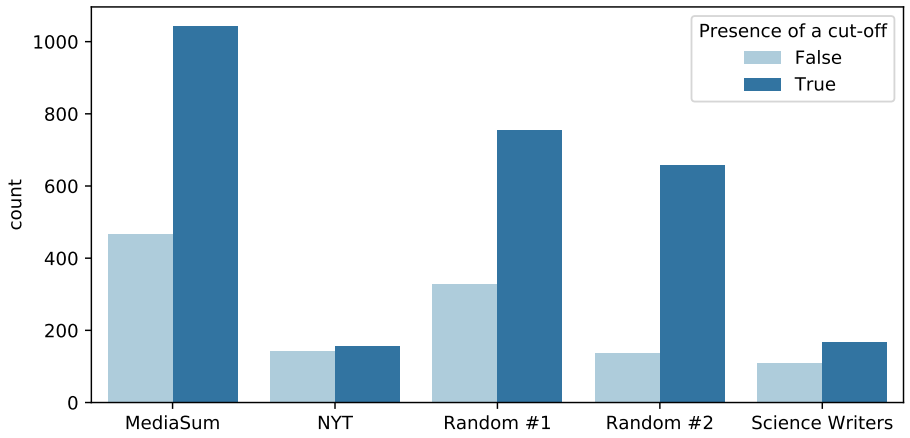


Fig. 10: Amount of egos with and without a cut-off point.

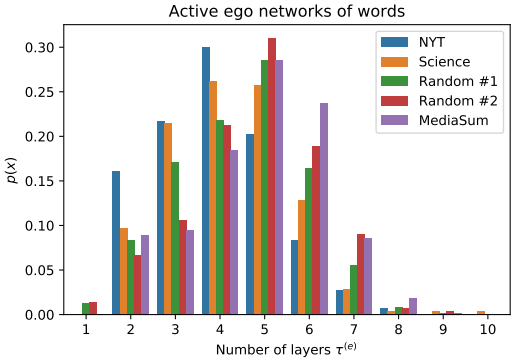


Fig. 11: Distribution of the number of circles for the active ego networks of words.

5 Results

The goal of this section is to fully validate the methodology proposed in Section 4. First, in Section 5.1 we show that the layered structure that was not present when considering the full ego network (Figure 5) emerges again when focusing on the active ego network, and we revisit its properties in Section 5.2. Then we evaluate the robustness of the methodology to a varying amount of input data (Section 5.3). Finally, we show that active ego networks are stable over time (Section 5.4).

5.1 Optimal circle size for the active ego network

We return to the initial motivation behind this work, namely the disappearance of the layered structure in the ego network of words within large textual corpora when failing to accurately identify the active portion of the ego network. This phenomenon was illustrated in Figure 5. By employing the methodology outlined in Section 4, we can now effectively isolate[5] the active component of the ego network and ascertain whether the layered structure reemerges. Figure 11 demonstrates that this is indeed the case. Comparing it with Figure 5, where the circles were computed on the full ego network, we observe that limiting the size of the ego network to the maximum number of actively used words shifts the mode from two circles to four or five circles, for all datasets. This means that the structure of the ego network fully emerges when the active part is properly isolated, similar to what happens for social ego networks. And that the methodology from Section 4 is able to properly identify the active part.

[5] It is important to note, as mentioned earlier, that we exclude all egos that have not yet reached their saturation point to ensure that the observed ego networks are mature and not partially empty.

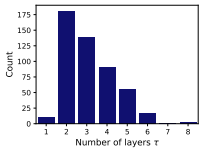
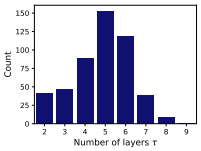
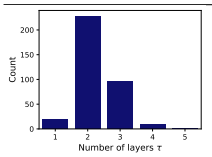
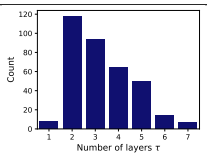
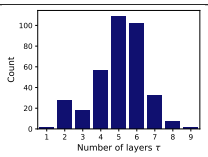
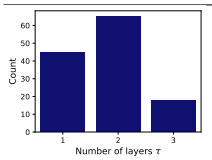
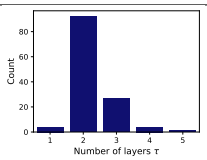
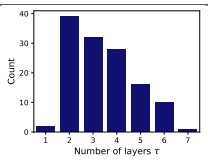
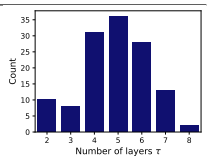
No cut	1st iteration	2nd iteration	3rd iteration
			
			
			

Table 4: Distribution of the optimal number of layers at each iteration of our recursive method on the Mediasum dataset. Each row contains egos with different numbers of total iterations, respectively 1, 2, and 3.

We now take a step further to demonstrate that the intermediate cut-off points achieved through the recursive method do not produce structured ego networks of words. In Table 4, we present the results for the Mediasum dataset exclusively, though readers interested in the results for other datasets can refer to Appendix A. This observed trend is consistent across all datasets. Each row in Table 4 corresponds to egos with the same number of total iterations (one iteration for the first row, two for the second row, and so on). The emergence of a structured ego network is indicated by the distribution of the optimal number of circles, shifting its mode away from the value 2 (which signals a substantial lack of structure) as the final iteration is reached.

When we consider the results from Figure 11 in conjunction with Table 4, we not only demonstrate that our proposed method automatically leads to well-structured ego networks by excluding "inactive" words but also establish that such well-structured ego networks only emerge at the conclusion of the recursive steps.

5.2 Revisiting the structural properties of the ego network of words

We can now investigate the properties of the active ego networks of words for the users in the datasets discussed in Section 3. Recall that egos that have not reached their cut-off point are excluded from the following analysis. The

remaining ego networks are reduced to their active size w^{n_a} obtained with the method of Section 4.4. From now on, we simplify the notation w^{n_a} to w.

The analysis in Figure 11 revealed that active ego networks typically consist of between 4 and 5 circles. It is worth noting that NYT journalists and science writers tend to have slightly fewer circles compared to random users and speakers in the MediaSum dataset. Notably, the ego networks of MediaSum speakers closely align with those of generic Twitter users #2. Interestingly, a similar optimal range of 4 to 5 circles was also observed in the social domain [12].

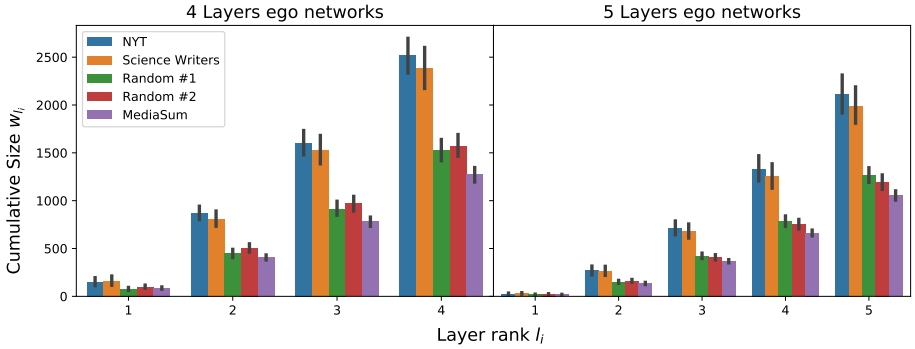
We now focus on the size of the ego network layers. For this analysis, we consider four- and five-layered ego networks, which are the most frequent cases in the five datasets, as shown in Figure 11, hence providing more samples for statistical reliability. In Figure 12a, the average layer sizes w_{l_i} are ranked from the innermost (l_1) to the outermost one (l_4 or l_5). Recall that the active size of an ego network, which corresponds to the total number of unique words before the cut-off, is also the size of the outermost layer. The layers of the ego networks from specialized Twitter datasets (NYT journalists and science writers) are on average bigger compared to random users and MediaSum speakers. Again, MediaSum speakers are quite well aligned with generic users on Twitter. According to the saturation curve methodology in Section 4.4, it means that they can handle a larger number of words before saturating their ability to bring new ones into their active vocabulary. The size of five-layered ego networks is consistently lower compared to the four-layered ones ($\sim$20% lower independently of the dataset). However, it seems that words have a similar distribution across the layers regardless of the dataset. We verify this property in the following.

We define the normalized layer size as the ratio between the layer size and the ego network size $\frac{w_{l_i}}{w}$. As can be seen in Figure 12b, normalized layer sizes are very similar across datasets. The penultimate layer $l_{\tau-1}$ consistently accounts for 60% of the ego network size, and the second to last layer $l_{\tau-2}$ accounts for 30%:

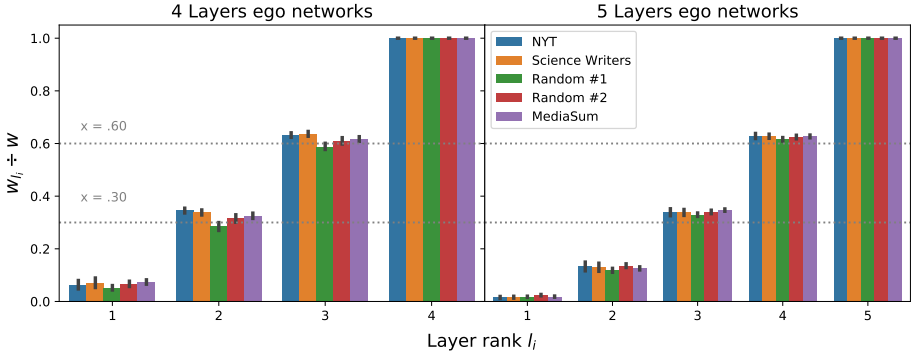
$$\begin{cases} \dfrac{w_{l_{\tau-1}}}{w} \simeq 0.6 \\ \dfrac{w_{l_{\tau-2}}}{w} \simeq 0.3 \end{cases} \tag{3}$$

We can observe the same pattern in the case of six-layered ego networks as well as for the penultimate layer of three-layered ego networks (Table 5). These values are very similar to those obtained in our previous paper [19] where the average ego network size was smaller. This means that the main difference between two ego networks with different numbers of layers is in the organisation of the inner layers. Note also that this regularity applies to all datasets, with no remarkable difference, further supporting the cross-domain generalizability of the ego network of words model.

The scaling ratio is a metric that describes how the layer size grows from a layer l_{i-1} to the outer layer l_i: $\dfrac{w_{l_i}}{w_{l_{i-1}}}$. As we can see in Figure 13 the ratio is very similar across the datasets for $i \geq 3$. The ratio tends to reach a value



(a) Plain



(b) Normalized

Fig. 12: Layer size

Dataset	# of layers	Layer Rank					
		1	2	3	4	5	6
NYT	3 layers	.16	.55	1			
	4 layers	.05	.32	.61	1		
	5 layers	.01	.11	.33	.62	1	
	6 layers	.00	.03	.15	.34	.63	1
Science Writers	3 layers	.25	.58	1			
	4 layers	.06	.33	.62	1		
	5 layers	.01	.14	.33	.63	1	
	6 layers	.01	.03	.15	.34	.63	1
Random #1	3 layers	.11	.53	1			
	4 layers	.04	.24	.58	1		
	5 layers	.01	.10	.30	.60	1	
	6 layers	.00	.03	.14	.33	.62	1
Random #2	3 layers	.13	.55	1			
	4 layers	.05	.26	.59	1		
	5 layers	.02	.12	.33	.63	1	
	6 layers	.00	.03	.12	.33	.61	1
MediaSum	3 layers	.15	.56	1			
	4 layers	.06	.31	.61	1		
	5 layers	.01	.11	.34	.63	1	
	6 layers	.00	.03	.14	.34	.63	1

Table 5: Average ratio between a layer size w_{l_i} and the active size of the ego network w , in all datasets.

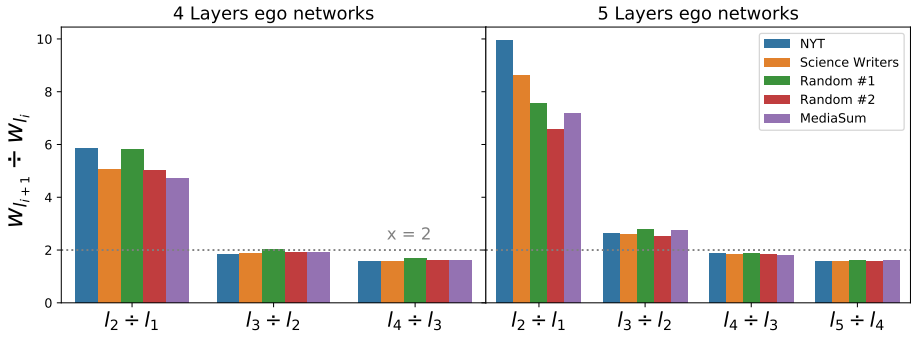


Fig. 13: Scaling ratio.

slightly below two toward the outermost layers. These results are the same as those obtained in the paper [19].

When comparing the current findings with previous research [19,20] that focused on ego networks of words, we must consider two aspects: first, the current work is based on more diverse and larger datasets, and second, the previous work did not specifically focus on the active network segment of the ego network (because a robust methodology for identifying it did not exist). Despite these considerations, the observations in the previous work [19,20] surprisingly align well with the current findings, particularly concerning the number of circles

(which were found to be between 5 and 7 in [20] vs 4-5 in this work) and the scaling ratio (approximately the same in [20]). However, when examining the absolute sizes of individual layers, we notice larger sizes in this work compared to [20]. To better understand this behavior, we can focus on the Twitter datasets, which are common to both studies (same users, shorter timelines in [20]). Both the similarities and differences in the ego networks can be explained by the fact that the observed timelines in [20] generally cover around or slightly less than the cut-off point. Consequently, the ego network structure becomes apparent, but some words are missing to make it fully complete (hence the smaller layers). Vice versa, the timelines we use in the current study cover much more than the cut-off point, hence, without a proper methodology to identify the active network, the resulting structure is meaningless (as shown in Section 4.3). Note that the slightly higher number of optimal circles in [20] can similarly be explained by an observation window below the cut-off point. While this may appear counterintuitive, the number of circles tend to grow as the number of data points decrease. This occurs because the clustering algorithm may detect spurious groupings when data points become more scattered.

5.3 Robustness of the methodology

In this section and the subsequent one, our primary focus lies on internally validating the proposed methodology for identifying the active network. We start with an analysys of the robustness of the methodology to the amount of available data. Specifically, the cut-off point of the active ego network should be a characteristic of each ego and not dependent on the size of the ego data fed to the algorithm. This implies that our algorithm should consistently determine the same cut-off point for a given ego, except when there is insufficient data to reach that point. In this section, we verify that this is the case.

Let us consider a tagged ego e whose saturation curve contains a cut-off point n_a. Recall that $\mathcal{T}^n \subseteq \mathcal{T}$ and $\mathcal{W}^n \subseteq \mathcal{W}$, for any $n < n^f$. When RecursiveCutOff in Algorithm 1 is fed $\mathcal{T}^n$ and $\mathcal{W}^n$ where $n < n^f$, it should return n_a if $n \geq n_a$ and n otherwise (if n is below the cut-off there is no cut-off to find). As n grows, then, the corresponding size of the active ego network will grow. When n reaches n_a, the active ego network is mature and should not grow anymore. This means that the active network size $\hat{w}^n$ for varying n should follow the ideal behavior :

$$\hat{w}^n = \begin{cases} w^n & \text{when} \quad n \in [0, n_a] \\ w^{n_a} & \text{when} \quad n \in [n_a, n_f] \end{cases} \tag{4}$$

In Fig 14 we plot the ratio $\frac{\hat{w}^n}{w^{n_a}}$. We expect $\frac{\hat{w}^n}{w^{n_a}}$ to grow from zero to one and then remain stable around one (implying that for any $n > n_a$, the calculated cut-off remains the same, regardless of the increasing size of the data being fed to the algorithm). Fig 14 confirms that the behavior of the calculated cut-off, and hence of the resulting size of the active network, is close to the ideal case in every dataset, despite some noise due to a lower number of ego networks in the NYT journalists and science writers datasets.

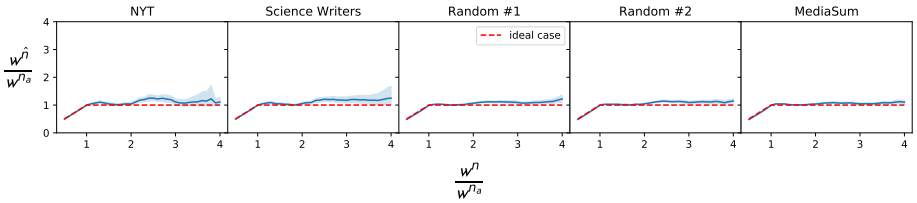


Fig. 14: The stability of the algorithm is close to the ideal case.

5.4 Temporal stability of the active network size

With the methodology introduced in Section 4, we are able to extract the active size of an ego network of words with respect to an observed tuple of tokens $\mathcal{T}$. This size corresponds to the volume of words actively used by the ego and whose boundary is associated with token t_{n_a} (from which the use of new words becomes rare). However, this count assumes that a word used at the beginning of $\mathcal{T}$ is still part of the active ego network. This raises the question of what would happen if we had started observing the language production of a speaker/user not from token t_0 but from a generic token t_δ. By shifting the start of the analysis from t_0 to t_δ, we study the dynamic evolution of the size of the active network, which is important because it allows us to assess whether the cognitive ability to add words to one's active vocabulary evolves over time.

To evaluate the temporal evolution of the active network size, we change the starting index of the sequence of tokens $\mathcal{T}^{n_f}$ from which we build the saturation curve. We call that shift δ, the updated tuple of tokens $\mathcal{T}^{\delta,n_f}$ and the corresponding word tuple $\mathcal{W}^{\delta,n_f}$. We build a new saturation curve, from which we extract an active network size w^{δ,n_a} (Figure 15). We want to compare w^{δ,n_a}, when δ varies, against the original active size w^{n_a}. If w^{δ,n_a} remains comparable to the second, it means that the active size of the network is stable over time. Thus, in the following, we study the ratio $\frac{w^{\delta,n_a}}{w^{n_a}}$. Note that the more we shift δ the more we run the risk of not observing egos for enough time and, consequently, of not having mature ego networks (much like the situation in which no cut-off could be found in Section 4.4). Thus, when shifting with δ we always make sure that, for each ego, at least n_a tokens are observed. This means that we operate in the range $\delta \in [0, \delta_{max}]$, with $\delta_{max} = n_f - n_a$. Note also that, differently from the previous section, here we never operate below the cut-off point n_a. In Figure 16, we choose a δ range from 0 to $5 \cdot 10^4$. That maximum was chosen because it is the largest value for which at least 25% of the ego network has a δ_{max} higher than it.

Following the above methodology, in Figure 16 we plot $\frac{w^{\delta,n_a}}{w^{n_a}}$ as a function of δ. We can observe that the ratio (hence, the size of the active ego network) remains stable when δ grows, independently of the dataset. This supports our hypothesis that the size and internal structure of the ego network are bound by

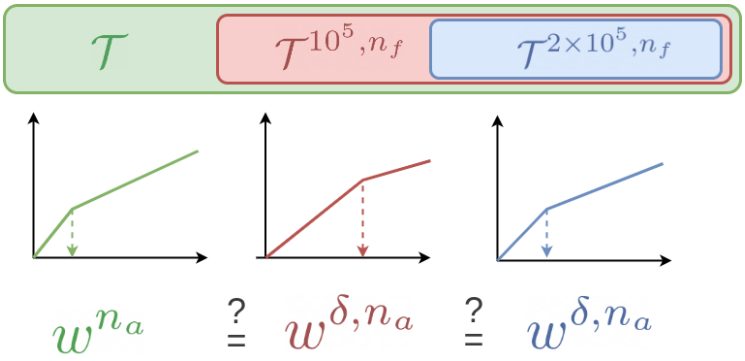


Fig. 15: The diagram illustrates the temporal analysis procedure of the active size of an ego network. A temporal change corresponds here to a change in the index δ of the first word of the sequence used to build the ego network. This change leads by construction to a different saturation curve from which we will extract and study the variability of the active part size w^{δ,n_a}.

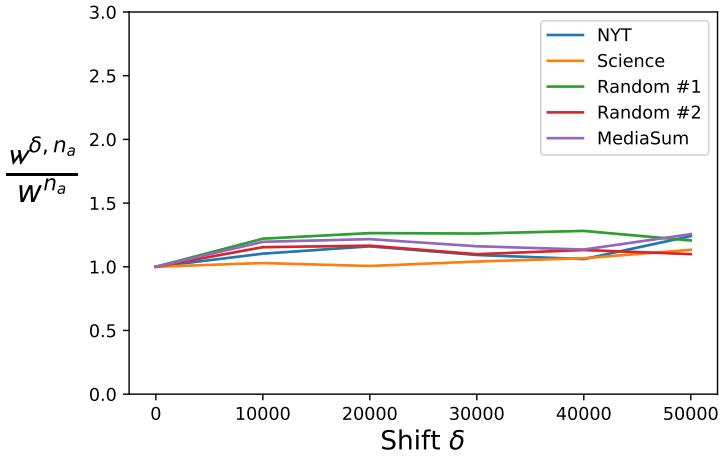


Fig. 16: The shift δ of the token sequence from which the ego networks are built has almost no influence on the active size w^{δ,n_a} on. In order to average that behaviour at the dataset level, we consider the ratio $\frac{w^{\delta,n_a}}{w^{n_a}}$ where the divisor is the original active size ($\delta = 0$). This ratio is consistently close to one (the maximum average value is 1.25, reached by the MediaSum dataset for $\delta = 5 \times 10^5$). These aggregated values are reliable since the 95% average confidence interval is only ± 0.08

cognitive constraints that are applied at different intensities depending on the individual, but which are themselves stable over time.

6 Conclusion

In this work, we investigated the cognitive limitations in human language production and presented the ego network of words as a model to capture structural properties associated with these constraints. The paper introduces the concept of an "active" part of the ego network, which represents the words actively used by an individual, and demonstrates that beyond this active part, the structure of the ego network becomes poorly organized. A robust methodology is proposed to extract the active part of the ego network, and its effectiveness is validated using interview transcripts and tweets datasets. Restricting our analysis to the active part of the ego networks, as commonly done when analyzing ego networks in the social domain, we have confirmed that the structural properties of the ego network of words, such as the number of circles and the scaling ratio between circles, are consistent across different domains. The presented methodology and findings have implications for various fields, including linguistics, cognitive science, and social network analysis. Future research can build upon these findings to explore additional aspects of language production and investigate the relationship between cognitive limitations and linguistic phenomena.

Acknowledgements. This work was partially supported by SoBigData.it. SoBigData.it receives funding from European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021. C. Boldrini was also supported by PNRR - M4C2 - Investimento 1.4, Centro Nazionale CN00000013 - "ICSC -National Centre for HPC, Big Data and Quantum Computing" - Spoke 6, funded by the European Commission under the NextGeneration EU programme.

References

1. Arnaboldi, V., Conti, M., La Gala, M., Passarella, A., Pezzoni, F.: Information diffusion in OSNs: the impact of nodes' sociality. In: Proceedings of the 29th Annual ACM Symposium on Applied Computing. pp. 616–621. ACM (2014)
2. Arnaboldi, V., Conti, M., Passarella, A., Dunbar, R.: Dynamics of personal social relationships in online social networks: a study on twitter. In: Proceedings of the first ACM conference on Online social networks. pp. 15–26 (2013)
3. Arnaboldi, V., Conti, M., Passarella, A., Pezzoni, F.: Analysis of ego network structure in online social networks. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. pp. 31–40. IEEE (2012)
4. Bentz, C., Ferrer Cancho, R.: Zipf's law of abbreviation as a language universal. In: Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics. pp. 1–4. University of Tübingen (2016)
5. Broadbent, D.E.: Word-frequency effect and response bias. Psychological review **74**(1), 1 (1967)

6. Brysbaert, M., Mandera, P., Keuleers, E.: The word frequency effect in word processing: An updated review. Current Directions in Psychological Science **27**(1), 45–50 (2018)
7. Caramazza, A.: How many levels of processing are there in lexical access? Cognitive neuropsychology **14**(1), 177–208 (1997)
8. Costa, A., Strijkers, K., Martin, C., Thierry, G.: The time course of word retrieval revealed by event-related brain potentials during overt speech. Proceedings of the National Academy of Sciences **106**(50), 21442–21446 (2009)
9. Davis, C.A., 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. pp. 273–274 (2016)
10. Dunbar, R.: The social brain hypothesis. Evolutionary Anthropology **9**(10), 178–190 (1998)
11. Dunbar, R.I.M., Sosis, R.: Optimising human community sizes. Evolution and human behavior : official journal of the Human Behavior and Evolution Society **39**(1), 106–111 (2018). https://doi.org/10.1016/j.evolhumbehav.2017.11.001
12. Dunbar, R.I., Arnaboldi, V., Conti, M., Passarella, A.: The structure of online social networks mirrors those in the offline world. Social networks **43**, 39–47 (2015)
13. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on information theory **21**(1), 32–40 (1975)
14. Hill, R.A., Dunbar, R.I.: Social network size in humans. Human nature **14**(1), 53–72 (2003)
15. Jenks, G.F.: Optimal data classification for choropleth maps. Department of Geographiy, University of Kansas Occasional Paper (1977)
16. Levelt, W.J., Roelofs, A., Meyer, A.S.: A theory of lexical access in speech production. Behavioral and brain sciences **22**(1), 1–38 (1999)
17. 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, pp. 281–297. Oakland, CA, USA (1967)
18. Mogotsi, I.: Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval: Cambridge university press, cambridge, england, 2008, 482 pp, isbn: 978-0-521-86571-5 (2010)
19. 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., Pedreschi, D. (eds.) Social Informatics. pp. 267–282. Springer International Publishing, Cham (2020)
20. Ollivier, K., Boldrini, C., Passarella, A., Conti, M.: Structural invariants and semantic fingerprints in the "ego network" of words. PLoS ONE 17(11): e0277182 (2022). https://doi.org/https://doi.org/10.1371/journal.pone.0277182
21. Qu, Q., Zhang, Q., Damian, M.F.: 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 **159**, 118–126 (2016)
22. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
23. Zhou, W.X., Sornette, D., Hill, R.a., Dunbar, R.I.M.: Discrete hierarchical organization of social group sizes. Proceedings. Biological sciences / The Royal Society **272**(1561), 439–444 (2005)
24. Zhu, C., Liu, Y., Mei, J., Zeng, M.: Mediasum: A large-scale media interview dataset for dialogue summarization. arXiv preprint arXiv:2103.06410 (2021)

A Appendix

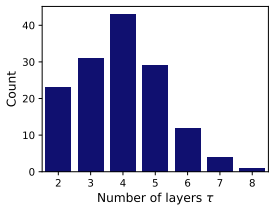
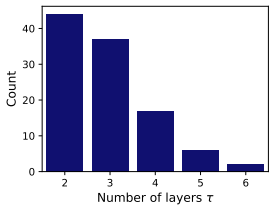
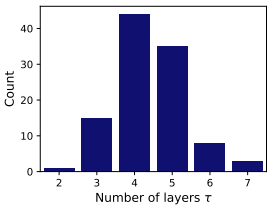
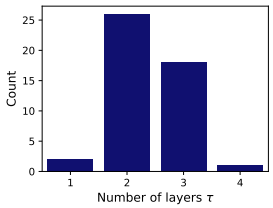
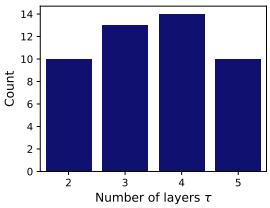
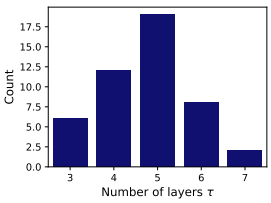
No cut	1st iteration	2nd iteration
		
		
		

Table 6: Distribution of the optimal number of layers at each iteration of our recursive method on the NYC dataset. Each row contains egos with different numbers of total iterations, respectively 0, 1, and 2.

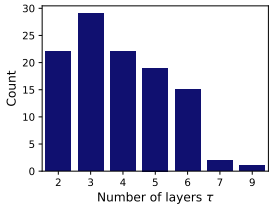
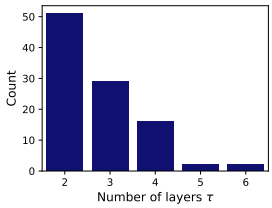
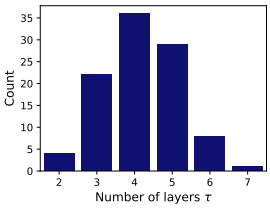
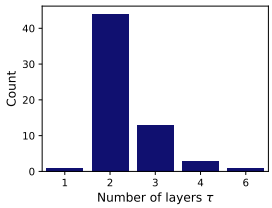
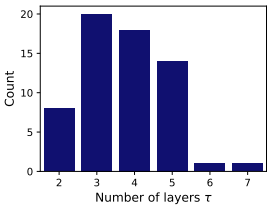
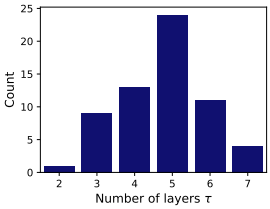
No cut	1st iteration	2nd iteration
		
		
		

Table 7: Distribution of the optimal number of layers at each iteration of our recursive method on the Science Writers dataset. Each row contains egos with different numbers of total iterations, respectively 0, 1, and 2.

Algorithm 1 Find the cut-off point of the saturation curve

Input: $\mathbf{t} = \{i : t_i \in \mathcal{T}^n\}$ and $\mathbf{w} = \{s(t_i) : t_i \in \mathcal{T}^n\}$, i.e. the datapoints of the saturation curve

Output: $\hat{n}$, i.e. the cut-off point.

$\hat{n} \leftarrow$ RecursiveCutOff$(\mathbf{t}, \mathbf{w})$

function RecursiveCutOff$(\mathbf{x}, \mathbf{y})$
  if IsLinear$(\mathbf{x}, \mathbf{y})$ **then**
    return last element of $\mathbf{x}$
  else
    $\hat{\mathbf{x}}, \hat{\mathbf{y}} \leftarrow$ SplitSaturationCurve$(\mathbf{x}, \mathbf{y})$
    return RecursiveCutOff$(\hat{\mathbf{x}}, \hat{\mathbf{y}})$
  end if
end function

function SplitSaturationCurve$(\mathbf{x}, \mathbf{y})$
  ▷ Subsetting notation "$[:n]$" means from first to n-th element
  ▷ "$[n:]$" means from n-th element to last
  $best\_n \leftarrow 1$
  $lowest\_rss \leftarrow +\infty$
  for $n = 1$ **to** $\max(y) - 1$ **do**
    ▷ get RSS from standard least-squares regression
    $rss_1 \leftarrow$ LinearFit$(\mathbf{x}[:n], \mathbf{y}[:n])$
    $rss_2 \leftarrow$ LinearFit$(\mathbf{x}[n+1:], \mathbf{y}[n+1:])$
    if $rss_1 + rss_2 < lowest\_rss$ **then**
      $lowest\_rss \leftarrow rss_1 + rss_2$
      $best\_n \leftarrow n$
    end if
  end for
  return $\mathbf{x}[:best\_n]$,$\mathbf{y}[:best\_n]$
end function

function IsLinear$(\mathbf{x}, \mathbf{y})$
  ▷ β_i is the Lasso coefficient associated with the polynomial term of degree i
  $\beta_1, \ldots, \beta_p \leftarrow$ LassoMaxVariableReduction$(\mathbf{x}, \mathbf{y})$
  if $\beta_1 \neq 0$ **then**
    return True
  else
    return False
  end if
end function

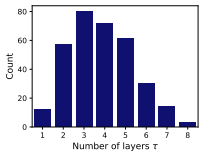
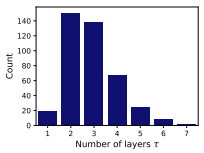
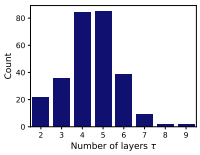
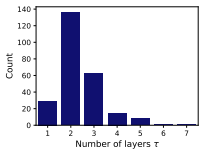
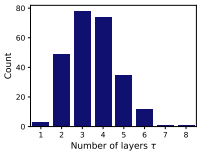
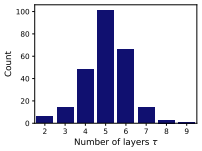
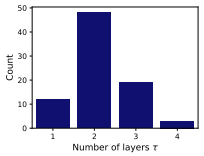
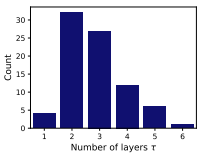
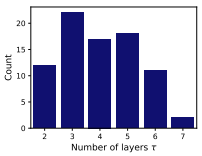
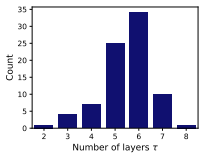
No cut	1st iteration	2nd iteration	3rd iteration
			
			
			
			

Table 8: Distribution of the optimal number of layers at each iteration of our recursive method on the Random #1 dataset. Each row contains egos with different numbers of total iterations, respectively 0, 1, 2, and 3.

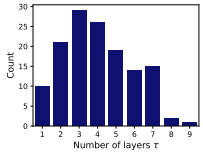
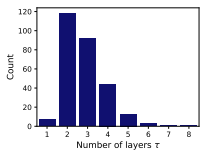
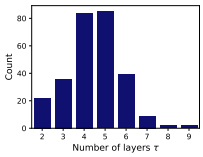
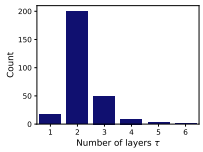
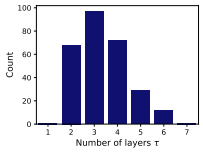
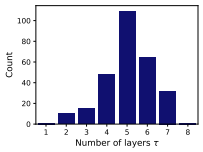
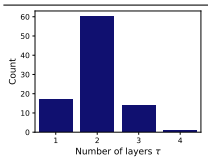
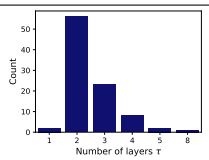
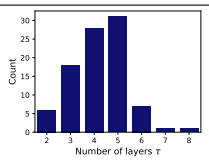
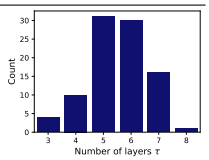
No cut	1st iteration	2nd iteration	3rd iteration
			
			
			
			

Table 9: Distribution of the optimal number of layers at each iteration of our recursive method on the Random #2 dataset. Each row contains egos with different numbers of total iterations, respectively 0, 1, 2, and 3.