SCUOLA
NORMALE
SUPERIORE

Classe di Scienze

Corso di perfezionamento in
Data Science

XXXIV ciclo

Using data science to uncover cognitive constraints in human behavior beyond social interactions.

Settore Scientifico Disciplinare **INF/01**

Candidato
dr. Kilian, Frédéric, Fabien OLLIVIER

Relatori

Dr. Chiara Boldrini

Dr. Andrea Passarella

Dr. Marco Conti

Dr. Felice Dell'Orletta

Anno accademico 2023/2024

"Si tu peux pas faire de grandes choses, fais de petites choses avec grandeur"

– Youssoupha, *Mon roi*

Acknowledgments

I would like to express my sincere gratitude to Dr. Chiara Boldrini and Dr. Andrea Passarella, for the trust they placed in me, their unwavering support (especially during the challenging months following the lockdown period in 2020) and their invaluable guidance from the inception to the completion of this research. Their expertise in data science and in research conduction significantly shaped my thesis. I am truly thankful for the time and dedication they invested in my academic journey, even beyond office hours after I transitioned to a full-time job in France.

I am deeply grateful to Dr. Marco Conti, the director of the IIT-CNR, for his scientific guidance and insightful feedbacks. I would like to thank Dr. Felice dell'Orletta for his precious advice in the field of natural language processing. I also would like to express my gratitude to Prof. Dino Pedreschi, the director of the joint Ph.D. program in Data Science, who provided annual precious feedback on my thesis. I would like to thank the panel members for their role in monitoring the advancement of my Ph.D. thesis: Dr. Francesco Marcelloni, Dr. Dominique Brunato and Dr. Stan Matwin.

I extend my sincere appreciation to the IIT-CNR for the material support (that included a scolarship, an office and free language courses), and to all members of the Ubiquitous Internet Research Group for their collaborative spirit and shared enthusiasm for research. I would like to thank the Scuola Normale Superiore which host the Ph.D. program, for providing financial support during more than three years, and giving access to its outstanding infrastructure.

In addition to the institutional acknowledgments, I would like to profoundly thank Agathe, my partner, for her patience and support, my family for the faith they had in me and their belief in my capabilities, and finally to all my colleagues and friends in Pisa who made this part of my life an unforgettable experience: Mustafa Toprak, Mattia Campana, Flavio di Martino, Pavlos Paraskevopoulos, Laura Renoldi, Vahid Zolfaghar and Marco Pallavicini.

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 the thesis, we postulate that similar regularities can be found in other cognitive processes, such as those involving language production. The thesis consists of three main parts.

In the first part, we leverage a methodology similar to the one used to uncover social cognitive constraints applied to the domain of language. More specifically, we are interested in understanding how individuals unconsciously structure their vocabulary. In order to investigate this claim, we analyse a dataset containing tweets of a heterogeneous group of Twitter users (regular users and professional writers). 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.

In the second part we carry out a semantic analysis of the model. 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 the innermost ring, which contains the most frequently used words, can be seen as the semantic fingerprint of the whole model.

In the third part, 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 (we used an extended version of the Twitter/X dataset and MediaSum, a preexisting dataset containing a large amount of interview transcripts). To address this, we define a methodology for extracting the active part of the ego network of words and validating it. The resulting ego network structures align substantially with the layer ego network of words obtained in previous chapters where only the active network was implicitly covered, confirming the model's robustness across different dataset sizes. 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 including spoken forms of communication.

Contents

1 Introduction

Language, as a complex cognitive activity, takes advantage of the unique capacity of the human brain for information processing but also embraces its limits. These limits are the central focus of this thesis. We demonstrate in the following chapters that language bears the marks of cognitive constraints, just like another human activity that has been the subject of in-depth study and has inspired our work: socialization. These two activities, like many others in everyday life, generate a growing volume of digital footprints, providing researchers with opportunities to better understand the specific nature of human behavior.

The human brain is an astonishing organ, capable of storing a volume of information several degrees of magnitude greater than any modern computer [179], and a powerful processing machine [97] that represents only 2% of body weight and yet costs 20% of metabolic load [12]. However, it is known to be subject to some limits, whether they come from the neuronal system [144] or from the optimization of energy consumption [95]. These cognitive limits shape the way we behave in daily life and how we make decisions. For example, our finite cognitive capacity is the reason for the existence of *cognitive biases* [86], which are the intellectual shortcuts we unconsciously take when we have too much or too little information available when making a decision. These cognitive constraints affect all areas of human activity, especially those that involve the processing of large amounts of information (driving a car, maintaining a conversation, learning a language, ...). They are the subject of a wide interdisciplinary field of study [113], which has been propelled by the development of the Internet and recent technological breakthroughs. In this thesis, we will investigate the evidence of such cognitive constraints using a data-driven approach. Indeed, the increasing interconnection between the offline and online worlds, known as "cyber-physical convergence" [31], opens up the possibility for public research institutes to study an ever-increasing range of human activities (such as social activities, sports, travel) on an unprecedented scale. In recent research work that stands at the frontier between psychology and anthropology and serves as a starting point for the thesis, Arnaboldi *et al.* [8] have been able to extend the understanding of human social skills by taking advantage of open-access data. Indeed, the authors extended the findings of R. Dunbar on the size of personal social networks [41] by using public interactions extracted from Twitter/X and

Facebook and provided significant results specific to online social relations. To this end, they developed an analysis tool based on Dunbar's work that focuses on individuals: the ego network model [40] (which we will introduce in the next paragraph and in detail in Section 3.1.3). The authors studied the characteristics of this model applied to users of Facebook and Twitter/X and identified structural invariants (common to all subjects of the experiment). These structural invariants were identified as "symptoms" of cognitive constraints that arise from human social behavior [8, 6, 5, 9, 7, 10].

In this thesis, we apply similar methods (using open-access data and an ego-centered model) to study language production. Specifically, relying on both written and spoken language traces from public sources, we build an ego-centered model adapted to the linguistic domain.

1.1 From cognitive constraints in social behaviours, to cognitive constraints in language production

Given the importance that social networks (both offline and online) play in our lives, in the related literature extensive attention has been devoted to the cognitive limits that affect how we entertain social relationships with each other. The cognitive efforts that we allocate to socialization have been extensively studied by anthropologists, and their findings [41] show that the social life of humans is constrained, through time and cognitive capacity, to 150 meaningful relationships per person (a limit that goes under the name of *Dunbar's number*, from the scientist who first postulated its existence). This limit is also observable in primates, where it is related to how many peers can be effectively groomed by animals to reinforce social bonds. In humans, these 150 social relationships can be grouped into classes of different intimacy. Specifically, anthropologists have found that the social relationships around the average individual can be grouped into at least 4 concentric layers [75, 184], starting from the innermost one which typically includes our closest family members. The typical sizes of these layers are 5, 15, 50, and 150. This model, organized in concentric circles and centered on a single individual, is called an *"ego-network"*. Just a few years ago, these constraints could only be studied via lab experiments, which are typically costly and time-consuming to arrange, and hence doomed to be restricted to small scales. However, with the advent of big data and online social networks (OSN), behavioural data are now available on a large scale and at a fine granularity. This opened up a new avenue of research that takes advantage of these data as a social microscope to better understand and characterize

human behavior. By exploiting data from public social interactions on OSN's, the "ego network" model was also confirmed for online relationships, with approximately the same structure [44] as for offline relations. The discovery of this structural invariant has represented a breakthrough moment in this research area. Many subsequent studies have leveraged this aggregate representation through social circles to better understand social-dependent human behaviour, such as how humans trust each other [164] or how they share resources and information [4, 5]. Socialization is just one of the many cognitive processes we experience in our daily life. Thus, it is reasonable to expect that similar limitations in our cognitive capacity yield characteristic structural properties in other domains as well.

Building upon the above considerations, in this thesis we set out to investigate the presence of an analogous structure and structural invariants in cognitive processes beyond the well-established social ones, especially in the domain of language production. Language is intimately linked to social skills, as there are hypotheses (that go under the name of *social gossip theory of language evolution* [42]) postulating that language has been developed as a more efficient way to groom social relationships: with vocal grooming, we can reach more peers at the same time. This human activity is subject to many cognitive processes that unfold transparently and exploit our cognitive abilities to their full extent (for example, during the selection of the lexical element that will symbolize the concept that must be expressed in sentence [96]). It is also subject to obvious cognitive limits, such as the persistence and volume of long-term memorization of the vocabulary. One of the most prominent examples comes from the observation by G. Zipf [188] in 1932 that the frequency of words in a corpus is inversely proportional to its position in the frequency table, which means, in other words, that we frequently use only a small set of the words available. It is also well known that our vocabulary size is limited: for example, an average 20-year-old native speaker of American English knows 42,000 words [18]. However, we also know that the human brain uses certain cognitive strategies to compensate for these limitations and achieve extraordinary results. 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 [96], thanks to the complex processing layers (semantic, syntactic, and lexical) involved in speech-related cognition [20]. The structure of the language is influenced by these cognitive strategies. For example, in most of the languages still existing, the most frequent words in a language are both the shortest [14] and the most quickly recovered ones in a speech production task [16, 138]. 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.

We conjectured that a structure similar to social "ego networks" may also be used to describe the way humans use words and that this structure can provide very significant information to characterize the peculiarities of individuals, similarly to the social dimension. Instead of classifying a person's social relationships in concentric circles, we decided to categorize the words they use in the same way, according to their frequency of use: the most used words in the innermost circles, and the least used words in the outermost circles. We called this model an "ego network of words". By using this model, we have been able to study the way in which individuals unconsciously organize their own vocabulary and to identify invariants that are likely to correspond to common limits of human cognition.

1.2 Organization of the thesis and main contributions

The thesis is organized as follows. Chapter 2 is a high-level introduction to the field of cognitive science. In Chapter 3, before exploring the work done on cognitive limits in language production, we focus on recent findings on the same subject, but in the context of another human activity, social relations. In Chapter 4, we present the data sets on which we worked, as well as the methodology for building ego network of words that is used throughout the thesis. We found that this model very well captures how individuals organize the words they use since we discovered that the optimal number of concentric circles (which classify each individual's vocabulary based on the frequency of use) is relatively stable in the studied population (between five and seven). We also found that the relative circle's sizes are extremely constant (*e.g.* the penultimate circle consistently accounts for 60% of the words in the ego word network). In Chapter 5 we carried out a semantic analysis of the ego networks of words obtained in the previous chapter. To this end, we performed a topic analysis of the words in each ring (a ring is the nonconcentric counterpart of a circle) and established a *"semantic profile"* for each ego network. We find that ring #1 (the innermost ring) has a special role in the model. It is semantically the most dissimilar and the most diverse among the rings. We show that there are a few important topics in that ring, but they also have the characteristic of being the predominant topics in the whole ego network. In this respect, ring #1 can be seen as the semantic fingerprint of the ego network of words. After focusing on the layer-based structure and its semantic properties, Chapter 6 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 defined a methodology for extracting the active part of the ego network of words and validated it using interview transcripts and tweets. The robustness of our method to varying input data sizes and temporal stability has also been demonstrated. The resulting ego network structures align substantially with the ego network of words obtained in previous work, where only the active network was implicitly covered, confirming the model's robustness across different dataset sizes. Moreover, the validation on the speech transcripts dataset (MediaSum) highlights the generalizability of the model across diverse domains and the ingrained cognitive constraints in language usage. Finally, in Chapter 7 we summarise and discuss the results obtained, then we introduce the research perspectives opened up by this thesis.

2 Background

2.1 Introduction to cognitive processes and cognitive constraints

2.1.1 What is a cognitive process?

A cognitive process is a high-level mental activity involved in acquiring, processing and retrieving knowledge. It is considered responsible for behaviour that can be expressed physically (*e.g.* speaking) or mentally (*e.g.* silent reading). Most cognitive tasks are a complex sequence of conscious and unconscious processes. When studying mental processes, the "cognitive level" also provides an abstraction of the underlying action of neurons (Figure 1). A cognitive process can be represented as a system that takes in a very large quantity of input, coming not only from outside (via the sensory organs), but also from information already acquired in a more or less conscious way (feelings, knowledge, beliefs) [175], and produces *"behavioural"* output (reading a book, chatting, playing, working, calculating, ...).

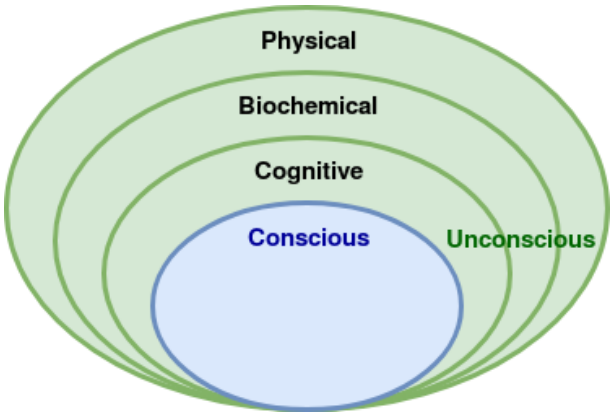


Figure 1: Venn diagram illustrating the different levels of mental processes, from the higher (conscious processes), to the lower (physical processes). [120].

Therefore, the cognitive system is a powerful tool for handling large amounts of information, with the purpose of choosing the best behaviour in a specific situation. It is generally broken down into functional sub-systems (with varying levels of brain activity) responsible for *"perception, memory, learning, emotion, intentionality, self-representation, rationality and decision making [...]"* [120] to which we can add language:

Perception is about the stimuli gathered by the body senses (vision, hearing, touch, taste, smell), but also more abstract information such as time perception. It also includes an initial interpretation phase necessary to transform this sensory signal into a mental representation that the brain can understand. This collection of information in enormous quantities relies on innate or acquired processing shortcuts that enable us to anticipate the presence of a signal or identify important information quickly and efficiently [175]. This kind of adaptation constitutes a set of biases which at some point provided a selective advantage. For example, it is due to this unconscious anticipation, coded in the neuronal circuit for the auditory perception of movement, that we can detect that an object is imminently approaching [119]. However, whether the information is auditory or visual, this perception can be fooled if the brain's presuppositions are wrong, for example when confronted with optical illusions that exploit the brain's ability to interpret perspective, as in the Müller-Lyer experiment [66].

Memory can be decomposed into different categories depending on their capacity and persistence, such as long-term memory, short-term memory and working memory [32] (the second and third are sometimes assimilated). The first is characterised by a large capacity (in terms of quantity of information) stored for months or years, with a slow forgetting decay. The second, on the other hand, is limited to a few seconds, with a much smaller capacity, of the order of 7 units plus or minus 2 [112]. Some theoretical frameworks distinguish working memory from short-time memory by its specific purpose of keeping a small set of information available for a task in progress (carrying over in digit addition, avoiding using the same ingredient twice in a recipe). Neuroimaging studies have shown that different types of memory activate different areas of the brain [165].

Learning is a broad research topic that is studied extensively by researchers in various fields such as psychology [24], neuroscience [178], science of education and development [178]. Unlike memorisation, which is often associated with a "declarative" piece of information that one can explicitly speak about and reuse later (*e.g.* memorizing a lesson), learning is more broadly associated with the acquisition of skills and attitude that will influence our behaviour (*e.g.* learning how to ride a bike).

Emotion have long been dissociated from cognitive processes, based on the

common assumption that there are brain areas dedicated to cognition, such as the prefrontal cortex, and those dedicated to emotions, such as the amygdala [132]). These latter areas are often considered to be evolutionarily fixed or *"primitive"*, and their actions are largely unconscious or obscure, such as the irrational fear of spiders and snakes [3]. However, recent studies show that this distinction is questionable and that both types of areas are responsible for so-called cognitive and emotional processes, such as the prefrontal cortex which, in addition to its cognitive role, plays a key role in affect [118] and emotion control [122]. Some studies tend to show, in the opposite direction, that emotions can help in rational decision-making [34, 52].

Intentionality is an aspect of consciousness that precedes a voluntary act by urging an effort in a specific direction. [99]. In [91], intentional actions are defined as opposed to stimulus-based actions, which are qualified as "external". In the same paper, the authors show that the same action (*e.g.* press a button) does not activate the same area of the brain depending on whether it is initiated by an internal will or whether it is a response to an external stimulus. Intentional actions require higher-level mental processes that can decide when and if an action should be carried out [91].

Self-representation is a concept that helps differentiate things that are linked to the self from the others, the former category receiving special attention from the brain [81]. Some papers are pointing out evidence of increased activity in the ventromedial prefrontal cortex when the brain is processing self-related content [46]. Researchers traditionally divide self-representation as a subject (controlling behaviour and receiving sensory information) from self-representation as an object of knowledge [57].

Rationality has long been described as a logical or even mathematical implementation that allows the human mind to reason [93, 101]. However, it has been recognised over the last fifty years that the brain's inference is far from complying with logical, statistical and probabilistic standards [103]. It has been shown that human rationality is heavily biased and relies on numerous error-prone cognitive shortcuts [23]. The notion of bounded rationality comes from the fact that the human brain has to make choices within a limited framework, in terms of the amount of information available (including that coming from memory), time, and calculation power [168].

Decision making is the fundamental high-level cognitive process by which choices are made. It involves an initial judgement phase, assessing the available alternatives [64]. As mentioned in the previous point, this judgement is subject to heuristic biases which contribute to reduce the choice complexity [85]. Decisions are based on both external events and the outcome of past experiences, which is why specialists think of this process as a feedback-dependent learning loop [63].

2.1.2 Known limits of the human brain

The human brain is able to store an almost infinite amount of knowledge and efficiently carry out highly complex operations, thanks to a parallel processing system spread over hundreds of trillions of neuronal connections [124]. However, in everyday life, it faces obvious limitations. For example, in the domain of visual processing, we often experience great difficulties in reacting to two simultaneous (or closely spaced) events [45] which is due to several information bottlenecks, including working memory [108]. In the middle of the 20^{th} century, psychologist George Miller leveraged Claude Shannon's theory of information [112] to better understand such cognitive limitations. He illustrates this with an experimental task in which an observer has to recognize objects (by multiple means of perception). Millers draws a parallel between the observer and the "communication channel" from the theory of information, where the input information is the real number of objects, and the output information is the number of objects recognized. The intuition of Miller is as follows: the more items there are to recognize, the less accurate the observations will be, as suggested by the theory of information which states that the quantity of information that can pass through the communication channel has a limit called "channel capacity". Some experiments have been conducted to quantify this limit [135], using stimuli that affect hearing, taste and vision. The results (Figure 2) show that the channel capacity remains low, independently of the kind of stimulus. In short terms, it means that humans are able to recognize on average 6.5 stimuli (with a low standard deviation). This number also corresponds to the number of points humans are able to count on average in a few milliseconds [90] (beyond this limit, the cognition relies on estimations and errors are made). Both experiments uncover strong limitations in information processing and attention.

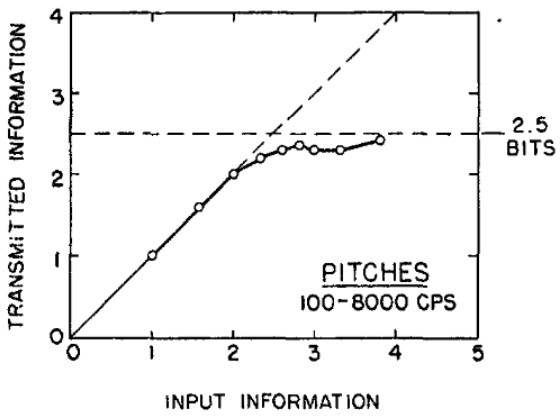


Figure 2: Experiment from Pollack [135] measuring the amount of information that human cognition is capable of transmitting. The observer listens to x sound samples with a different pitch and recognises y (those numbers are expressed in bits *i.e.* $log_2(x)$ *and* $log_2(y)$). In this experiment, y reaches a limit that corresponds to a channel capacity of 2.5 bits (around 5.6 sounds).

Using information theory to better understand cognitive costs

In a recent paper, Zénon *et al.* argue that the cost of a cognitive process can be viewed as the cost of modifying a mental state [182]. In this theoretical framework, a mental state corresponds to prior knowledge in Bayesian statistics and can be written as a probability distribution $p(y)$ of a behaviour y. Consequently, any task involving cognitive control must update this distribution $p(y|x)$ to choose the best behaviour for a given environment x. Zénon *et al.* [182] take as an example an experimental task where the subject has to press one out of four numbered buttons. In the absence of additional information, the prior knowledge $p(y)$ is equally distributed over the four available actions ($y = 1, \ldots, y = 4$). When a screen shows the number of the button to press, the knowledge priority is updated so that $p(y) = 1$ for the correct button. The more distant $p(y)$ is from $p(y|x)$, the greater the effort required to update it. This cognitive cost is estimated using the Kulback Leibler divergence measure $KL(p(y), p(y|x))$. It has been shown in some experiments that reaction time is linear with this value [74, 80]. The paper goes further by dissociating the environment y from its mental representation y', and by adding a context variable, but the aim remains the same: estimating a cognitive cost based on successive updates of internal beliefs. This work, based on information theory, provides a framework to

explain why some simple or dual tasks reach a cognitive limit. Most notably, it predicts that: *"certain kinds of tasks - which have many degrees of freedom, are unfamiliar, necessitate to go against natural biases, have variable natural structure or low signal to noise ratio - will lead to large information costs."*

Many degrees of freedom: tasks with many degrees of freedom, such as creative tasks, have a large number of responses y over which the distribution $p(y)$ is widely spread. Any update $p(y|x)$ of this prior knowledge (for example, choosing a series of moves from thousands possible in chess) will therefore be costly.

Unfamiliar tasks: for unfamiliar tasks, the best prior distribution y (or its encoding) is unknown. Consequently, any update is likely to completely change this distribution, and therefore to be costly.

Counteracting prior policies: corresponds, to adapting to a new environment, a new language, or new interlocutors. Following the same reasoning as above, if we consider that the previous distribution of y is no longer valid at all, then the cognitive cost, which is proportional to the amount of change, is high.

Task switching and dual tasks: It is widely accepted that carrying out two tasks at the same time (listening to someone and reading a book) or in sequence (working between two phone calls) is very difficult and tiring. To take this parameter into account, the framework uses the concept of context (also expressed as a probability distribution). During the execution of a task, this distribution specialises in optimising task execution. For example, it may correspond to the acquisition of useful information that makes the task easier and easier to perform. Any change of task therefore begins with an unsuitable context probability distribution (the more time spent on the previous task, the more specialised this distribution, the more difficult it is). This helps to explain the inertia of task efficiency, and why performing several tasks in the same short span of time involves a significant cognitive cost.

Performance/information rate: A final aspect that affects the cognitive cost of a task concerns the clarity of the stimuli to which one should respond. This is simply the signal-to-noise ratio of the environment variable x. A poor ratio leads to poor compression of x, a more spread-out distribution, and, finally, a higher cognitive cost for updating $p(y)$ [182].

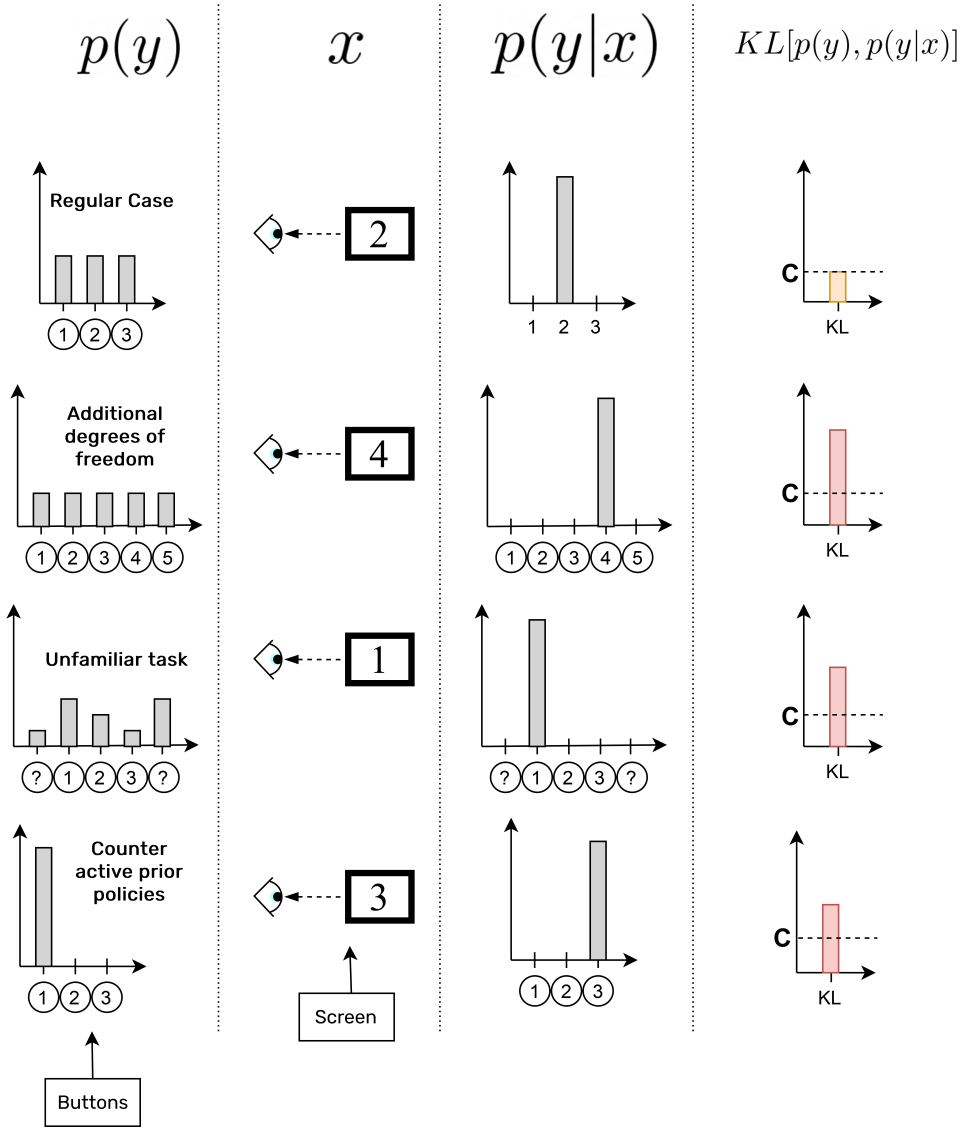


Figure 3: Toy example illustrating the update (column 3) of the prior knowledge (column 1) following instructions on a screen (column 2). Knowledge corresponds to the distribution of probability over buttons to push. Kullback-Leibler divergence is a measure of distance between prior and updated knowledge (column 4). We call C that distance for the regular case (for a purpose of comparison).

This concept of optimal distribution is in line with studies showing that the brain adaptively encodes information according to its redundancy in the environment: information that is extracted or used frequently should entail a lower cognitive cost than more scarce information [49, 29]. In the language domain, we can make a similar observation: the most frequent words are also the shortest for the majority of languages. [14].

Biological aspects of cognitive costs

An open question concerns the benefit to the brain of avoiding these costs and their origin. Do they come from structural limits in the brain's information processing or from an energy cap? The finite number of neurons and their speed of communication impose necessary restrictions on the amount of information that can be processed [144], any intensive use of an area of the brain by one task would lead to a reduction in the efficiency of other tasks using that same area, including vital tasks. Also, as it has been shown that energy consumption in the brain does not vary [95, 153], this may induce indirect competition even for distinct areas of the brain for the main resource, glucose (useful for glycolysis reaction). It also seems that energy consumption is partly due to structural modification of information induced by the repetition of certain tasks [172] (which corresponds to the update of prior belief mentioned earlier in Section 2.1.2). However, even if this immediate alteration involves a significant cost, it is profitable in the long term because the task will then be less costly in terms of energy resources [154]. In the case of language learning, studies based on fMRI have demonstrated this adaptation through repetition [181]. In the experiment, repeated exposure to unfamiliar syntactic structures led to increased cerebral activity in specific areas of the brain (left inferior frontal gyrus, medium and superior gyrus) compared with repetition of structures that were already known (mother tongue) which, on the contrary, saw progressively reduced activation. This difference suggests that repeated activation allows the construction of a mental representation linked to a new language feature [68]. This study highlights the cognitive effort involved in iterative learning, in particular that of a language to create a neural structure that will ultimately enable the signal to be processed more efficiently. Interestingly, learning a syntactic structure that is different from, but still close to, a known structure (e.g. switching from one Romance language to another) also produces decreasing activation (albeit not as strongly as with the mother tongue). This means that a similar pre-existing mental representation can be used to process the syntactic structure, and only minimal adaptation is required.

2.1.3 Bypassing cognitive limits: heuristics and cognitive shortcuts

The brain limits introduced in Section 2.1.2 strongly impact decision-making. When making a choice, one should identify and compare all the possibilities, which require too many cognitive resources. "Bounded rationality" refers to the fact that decisions have to be made with a limited quantity of information and time [149, 150]. The brain operates some effort-reduction techniques which makes decision-making tractable [147]. Cognitive biases arise from these limitations. In the right context, they are useful to achieve an action more quickly, but sometimes they lead us to make irrational choices. Cognitive biases are generally classified in four groups [13]: those which tackle the mass of input information (*eg.* confirmation bias [180]), those which infer meaning out of chaotic data (*eg.* survivorship bias [177]), those which accelerate the cognitive process (*eg.* less is better effect [76]), and those which strengthen memorization (*eg.* misinformation effects [11]). These shortcuts alter our perception of the world and can be exploited for manipulation purposes, for example by the advertisement industry.

In their paper [73] Hertwig and Todd argue that cognitive limits enable important cognitive functions:

- Limited knowledge and memory activates a better recognition of important ideas [151, 60]. It relies on the fact that what has been learnt should be more important than what has been forgotten.
- The human brain tends naturally to interpret his environment with causal relations [128]. These correlations rely generally on a very small sample of empirical data: as explained in Section 2.1.1 the working memory is limited to a few items. This limitation can be an asset since a small sample allows better early detection of correlations (when they exist) [89].
- A small working memory could be also an advantage in language acquisition. Analyzing fewer words at the same time can prevent learners from making wrong generalizations [47]. Enlarging step by step the working memory could be a winning strategy, and be one of the many reasons why babies have a steep language learning curve.

In the same paper [73], Hertwig and Todd question the fact that limited cognitive capacity is the sole responsible for simple heuristics and cognitive shortcuts. For them, if more complex heuristics had been a selective advantage, cognitive limits would have been "pushed" further by extending the

brain volume. They make the hypothesis of a reverse causal relation: the speed and reliability of simple heuristics have been a selective advantage and there is no need for larger cognitive capacities which could represent a waste of energy. However, this hypothesis has yet to be proved.

2.2 Studying human cognition in the age of big data

2.2.1 A historical perspective of cognitive science

A "cognitive revolution" came in reaction to the movement of behaviourism in the middle of the 20th century. The latter approach considers that mind studies should ignore non-observable mental processes and focus on behaviour (*e.g.* the experiments of I. Pavlov on conditioned reflexes). From this perspective, the behaviour is the sole result of the interactions of the subject with the environment. The main objective of behaviourism was to prevent any pseudoscientific interpretation of the unobservable phenomenon that unfolds between the external stimulus and the body reaction [152]. However, some researchers like N. Chomsky argue that some behaviours like language are a product of innate and complex cognitive mechanisms He and many other researchers felt the necessity to break down the barriers between several disciplines to let new ideas and mind theories emerge (Figure 4). These disciplines are: psychology, philosophy, neuroscience, computer science, anthropology and linguistics. G. Miller predicted in 2003 that each of the links that connect the six disciplines could potentially give birth to a whole field of research (such as computational linguistics).

2.2.2 The era of big data

The big data era is characterized not only by the high quantity of data provided by human activity but also by the growing variety of situations where we generate digital traces. The result is that the "cyber world" is increasingly shaped by data coming from the "real world", both at an individual level (*e.g.* a friend or restaurant recommendation based on location) and at a global scale (*e.g.* Twitter/X volume of tweets during an important event). The opposite is also true: the physical world is now influenced by the use of websites and apps embedded in smartphones that we bring with us everywhere. Both worlds are interacting and tend to overlap: this phenomenon is called cyber-physical convergence [31]. New types of datasets have emerged, providing researchers with valuable material for cognitive research. They can be records of *physical* or *online* activities, obtained *naturally* or in an *experimental* context (Table 1).

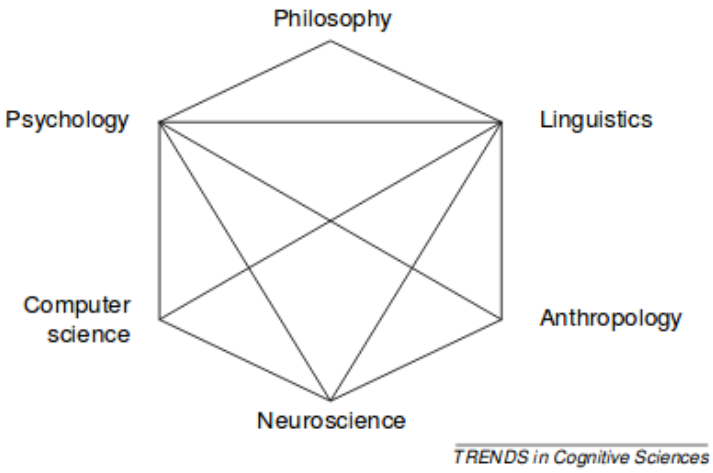


Figure 4: Revolution of cognitive science is an interdisciplinary approach [113]

Records of physical world events: A big part of the data is collected in order to deliver context-aware services, more convenient to the end user. For instance, wearable technologies are designed not to be invasive, such that information recorded is not perturbed by the device itself.

Records of online events: Online events are collected when users are browsing apps or web pages. They are traces of a behaviour that is more "internal". This kind of information is a big asset for marketing purposes [30] because it can, for instance, give a better understanding of a whole consumer buying process. The engagement of users is often modeled as a funnel: the buying process is a succession of steps that a decreasing number of people are reaching. Behaviour analysis can help identify what makes the user abandon the purchase, and maximise the number of people who reach the final step. It is even easier to study behaviours directly via online social networks because information is mostly shared voluntarily (geo-tagged photos, messages, reactions to a post, ...). Many research projects are using social media to study various aspects of human science, including sociology [44] and psychology [94].

Naturally occurring datasets: In the past, cognitive scientists have

	Online Events	Offline Events
NODS	Social network content	Wearables records
Experiments	Mechanical turk experiments	Classic cognitive experiment

Table 1: Example of different data sources for cognitive studies.

heavily relied on, typically small-scale, lab experiments. With the advent of the big data era, the data sources that they can leverage have multiplied, as well as the volume of data. In this context, naturally occurring datasets (NODS) which are typically produced for non-scientific purposes, reveal natural behaviours in an uncontrolled context. One of the intrinsic advantages of NODS is that they avoid experiment-related biases [61] that could alter one's behaviour. Their volume is much bigger than experimental datasets.

Datasets from crowdsourced experiments: Lab experiments can themselves be enhanced by internet technologies. Stewart *et al.* estimate that half of the cognitive science experiments will involve a crowdsourcing platform such as Amazon Mechanical Turk [159]. The latter service allows its users to buy workforce at a large scale in order to complete a given task. For example, it can be used to collect manually labeled data for feeding a machine-learning algorithm. Cognitive scientists can take advantage of this opportunity to create online experiments, measure reaction times with good accuracy, and collect results qualitatively as good as those obtained in labs [157]. Although, even if the population that takes part in the experiments is larger and more diverse than before (with respect to onsite experiments), it is still not representative of the general population. For example, Amazon Mechanical Turk workers (the population that completes the tasks) are more likely to be unemployed, American or Indian, young, and liberal [22]. Stewart *et al.* argue that even if this method cannot solve alone the "crisis of reproducibility", reducing the time, effort, and cost needed for data collection allows researchers to test a bigger amount of ideas [159].

2.2.3 From predictive models to valuable cognitive insights

Understanding online behaviours can serve many research and commercial purposes. In his manifesto for a new cognitive revolution, T. Griffith remarks that most of the studies only rely on a shallow understanding of these behaviours [67]. Instead of exploring the mind and the cognitive mechanisms, the goal is generally to obtain the best ratio of good predictions about users' incoming actions. For example, most of the recommendation systems rely on collaborative filtering which is a very easy and efficient way to predict one's choice based on the previous choices of similar persons [140]. The result is often right, but it offers no explanations about the reasons for that choice. Interestingly Griffith makes a parallel between this approach and behaviourism: our past actions and those of similar people are sufficient to predict our future actions [67]. Then he exhorts for a new cognitive revolution that also considers the existence of a mind, with internal states and innate mechanisms. For Schulz *et al.* [145], who studied the decision-making process for the selection of online food delivery service, this would not only improve our knowledge of the human brain, but also the performances of recommendation systems.

3 Related work and motivation

3.1 Cognitive constraints in social relations

3.1.1 The social brain hypothesis

The large size of the human brain is related to the high number of social relationships we establish during our lives: that is the hypothesis Robin Dunbar proposed in 1998 [43]. He states that the human brain can be approximately divided into three parts which developed at different periods of our evolution: the reptilian brain, the mammalian brain, and the primate brain [106]. The latter, which is also known as the neocortex, increased significantly in the primate group [127] (in fact, it accounts for 50% to 80% of the brain volume). Thus, studying the role of the neocortex is the key to understanding the causes of the human brain size. The neocortex is considered responsible for high cognitive perception, language, and consciousness. Hence, the neocortex holds the main assets for having complex relations with other individuals. This is the reason why Dunbar conjectured that the brain size (more precisely the visual cortex in the neocortex) could be correlated with the number of relationships species can maintain throughout their lives. In his studies, the number of relationships is measured as the mean group size within each species. Group sizes are crucial because they need to be big enough to defend against predators, but small enough to avoid short-term conflicts and sharing resources [163]. The correlation between neocortex size and group size is very strong among prosimians, simians and humanoid species. Recent studies based on neuroimages provided more credit to this hypothesis [98, 136, 87]. Regression performed on this data predicts a mean group size of 150 for humans (according to the neocortex size). This number has been called "Dunbar number" afterwards.

3.1.2 Human social skills: a neuroscientific perspective

The study of social behaviour follows two main different approaches: evolutionary psychology and sociobiology, which is an extension of ethology dedicated to animals, and social psychology, which considers human behaviour as an independent object of study. R. Adolphs [2] suggests that neurosciences can bridge the two approaches by considering one part of cognitive processes as innate and instinctive (like for animals), and another part which is acquired, more conscious, and is a specificity of humankind. It is therefore a non-monolithic and complex field of study, requiring the study of innate and acquired processing of social information. For example,

one of the skills that seems specifically human is the ability to attribute emotions to another individual, a particular mental state or beliefs. This is known as the theory of mind [137, 148], and is acquired in humans around the age of 4. This ability seems to be linked to the frontal lobe [161] and, more importantly, the prefrontal cortex [51, 56], which are both part of the neocortex, the area of the brain used by R. Dunbar to predict the social capacity [43]. Studies have shown that this area is activated when a subject performs theory of mind tasks (*e.g.* detecting a social 'faux pas' [160], appreciating humour [59], and viewing erotic stimuli [88]). However, researchers also highlighted the role of the amygdala in the theory of mind, which seems to be responsible for decoding facial expressions [19]. While the amygdala is known to be responsible for 'instinctive' behaviours such as fear or threat perception, the prefrontal cortex processes more 'self-regulatory' information which belongs to the second category of cognitive processes (social and moral judgement) [117]. This illustrates the simultaneous involvement and exchange of information between these two types of zones.

3.1.3 The ego network model

Definition of an ego network

In order to explore the complexity of social relations, we can rely on the graph-theoretical ego network model. It is a micro view of a social network centered on one individual. A social network connects individuals who interact with each other. The ego network focuses on a single individual, denoted *ego*, and the people interacting with her, called *alters* [163]. We rely on the ego network definition that does not consider the links between the alters. As humans do not spend their social effort equally among their relationships [143], we can cluster the alters having a similar ego-alter relation strength. It has been found that the number of these clusters (known as *circles*) is often close to 4-5 and that their size forms a recurrent pattern of 5-15-50-150. Hence the scaling ratio between adjacent circles is surprisingly stable (around 3) [185]. Zhou *et al.* point out that these numbers (they call this pattern "discrete scale invariance") can be found in several human organizations such as Land Army Corps [185]. They argue that it could be a discrete hierarchy deeply rooted in the human mind corresponding to an artifact of a "herding behavior". The first circle, which is called the "support clique", is composed of the five most reliable ego's relatives and friends [7] (Figure 5). The second circle ("sympathy group") is composed of 15 alters (including the support clique) whom the ego contacts at least once a month,

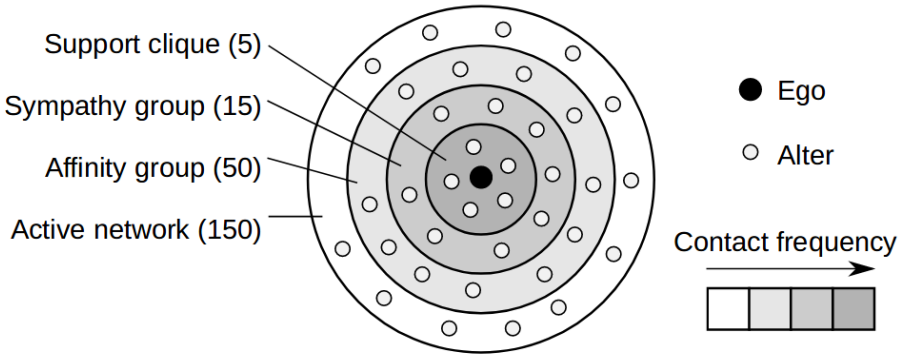


Figure 5: The ego network model (figure from [6]).

the third is the "affinity group" and the last is the "active network". The latter circle is composed of about 150 alters, contacted by the ego at least once a year.

A characteristic fingerprint of these social circles is their scaling ratio (i.e. the ratio between the sizes of consecutive layers), which has been found to be approximately around 3, regardless of the specific social network considered. Interestingly, both real-life and online social networks follow this social organization [44, 70, 115, 62]. The discovery of this social structure (stratified in concentric layers) and its invariants (the number of layers and scaling ratio across different and heterogeneous social networks) has represented a breakthrough moment in this research area. Many subsequent studies have leveraged this aggregate representation through social circles to better understand social-dependent human behaviour, such as how humans trust each other [164] or how they share resources and information [4, 5].

Ego network structures across different social interactions means

While the layered social circles structure was initially discovered in offline (*i.e.* real life) social relationships, this hypothesis has been later confirmed for a variety of social interaction means. Mac Carron *et al.* also identified distinct layers in ego networks coming from a mobile phone call dataset [105]. They show that the scaling factors between consecutive layers are close to 3. Generally, the active network size is smaller than Dunbar's number, but a hypothesis is that phone interactions are just a fraction of all social interactions.

Online social networks (OSN) introduce new means of "virtual" interac-

tion with other people (messages, likes, photos, invitations, polls, groups ...). Since they are really popular, we can study social phenomena at a very large scale (*e.g.* world scale for Twitter/X and Facebook). They also contribute to a global cyber-physical convergence: online actions affect the real world, and vice-versa, even at a very local scale. Geolocated posts, Facebook events, and Facebook safety checks are examples of cyber-physical convergent tools. It means that our online social behavior tends to reflect our real offline attitude. Moreover, many OSNs give the public access to massive amounts of their data. This data can take many forms: text, time/space coordinates, social temporal, weighted/directed graph with labeled nodes, bipartite graph (event participation), and time series (tweet writing frequency). Even if communicating through an online social network allows us to contact everybody in the world in a few seconds, it seems that our online social ego networks are still ruled by the same constraints as in the offline world [44]. In this study, the authors built ego networks from Facebook and Twitter/X datasets. The tie strength between individuals can be measured as the frequency of public posts or tweets sent from one to the other. After performing an alters cluster analysis for each ego, they highlighted that the most common case for egos is to have 4 clusters (Facebook) and 5 clusters (Twitter/X) where clusters represent the social circles of the ego networks. They also found that the scaling factors between consecutive layers is close to 3. With respect to the size of the layers, they also found a pattern close to 5, 15, and 45. In addition, they underline two differences: online social ego networks possess an inner circle of size 1.5 and the outermost layer is smaller than 150. The former is actually a validation of a long-standing hypothesis in anthropology that postulated the existence of an additional circle inside the support clique. The small-scale data from lab experiments was not sufficient to validate this hypothesis, but OSN data is. This result shows that OSNs can be used as a social microscope to study human interactions at a very large scale. The second finding (the size of the outermost layer smaller than 150) was also a surprising discovery. OSNs make communication easier than ever: we can access them through a smartphone from everywhere, contact someone in a few seconds thanks to the best user experience developments, receive communication incentives (recommendations, notifications), and connect and keep in touch with new people that we do not know offline. Despite being a medium that might facilitate interaction between distant people, this new communication channel does not change the "social capacity" of humans. The key idea here is that human online social behavior is not fundamentally different from the offline case. Therefore we can use the data gathered online for studying human behavior offline (where there is a serious lack of data).

Ego network structure and information diffusion

Patterns of information diffusion are strongly dependent on the ego networks properties [7]. The first evidence is that information will more likely flow from pairs that mutually trust themselves. This level of trust can be measured as the frequency of contact between two persons. Hence the different trust levels of an individual towards others can be modeled with its ego network circles (the innermost circle is the most trusted one) and each layer has a different impact on information diffusion. The structure of ego networks might significantly impact the properties of information diffusion. For example, assuming the diffusion of very trusted information, that is propagated only between the most intimate relationships (*e.g.* the support clique), it has been found that the number of hops needed by a piece of information to cross the network might be significantly higher than the famous concept of "6 degrees of separation" [7]. However, the outermost layers are also very important for high-scale information diffusion. If the ties that do not belong to an ego network (*i.e.* the weakest ones) are removed and not used to diffuse information, the information diffusion process still reaches about 97% of the nodes [7]. However, if the "active group" ring (*i.e.* excluding alters from inner circles), is removed from the dissemination process, then the information diffusion process reaches only about 30% of the nodes. It means that information must flow at least a few times through the weakest links in order to spread at a large scale. This evidence supports the hypothesis of Granovetter [65] that weakest ties allow inter-group communications.

Relationships turnaround

The fact that the cognitive capacity for social interaction is limited also impacts the turnaround of relationships in one's ego network. This has been analysed in several papers, using again data from OSN's [7]. New alters enter the online social ego network at a very high rate the first months then at a constant rate [183, 176, 176]. As the number of alters remains constant after a while [114, 176], it means that some alters also leave the ego network at a constant rate, so that the social effort of the ego remains unchanged. Arnaboldi *et al.* calculated the turnaround rate for each of the ego networks ring [7]: there is a higher turnover for sympathy group than for support clique (similarly to offline relationships, the more we are close to a person, the less likely she will be "replaced").

3.2 Cognitive constraints in language production

3.2.1 Language is more than a communication channel

It is increasingly acknowledged that language function is not limited to organizing the flow of information into and out of human cognitive systems (which is equivalent to seeing language as a mere channel). Carruthers [21] posited that language is crucial for communication between brain modules (modules are clusters of interconnected areas responsible for a specific brain function [170, 134, 155, 156]). When studying children's ability to activate modules responsible for different mental activities at the same time (*e.g.* sense of smell, shape and color recognition), it has been shown that having acquired the capacity of language resulted in significantly better performances [71]. In another experiment of the same type, researchers found that overloading the language module with a linguistic task significantly degraded the efficiency of other simultaneous cognitive tasks [72]. This suggests that language processing capacity is both limited and useful for module communication. We can also mention internal thoughts, which occupy a considerable amount of our time (some studies argue that we spend an average of 50% of our time having internal dialogue [78, 79]). Researchers believe that inner thoughts enhance our task-switching [48] and problem-solving performance [39, 104]. This is a further indication that language plays an active role in the internal information flow of the brain.

3.2.2 Language fits the brain structure

It is widely known that children have great capacities for learning words and speaking by mimicry. There is a consensus that the brain is fully adapted and optimized for this highly complex task of language learning. In a recent paper [27], Christiansen *et al.* suggest that the real question is not *"why is the brain so well adapted to learning language ?"* but *"why is language so well adapted to be learned by the brain ?"*. They posit that language has been subject to evolutionary pressure (in the Darwinian sense) in order to be assimilated as efficiently as possible by the brain. The authors clearly draw a parallel between the evolution of a language and that of a biological system: it is subject to random mutation (*e.g.* the emergence of new words), its characteristics are transmissible (through language learning), and it is subject to selective pressure linked to its environment, which is human cognition. It should be noted here that in the evolution process, a language is not compared with a biological individual, but with a whole species: a language is, in fact, a coherent system materialized by the activity of its

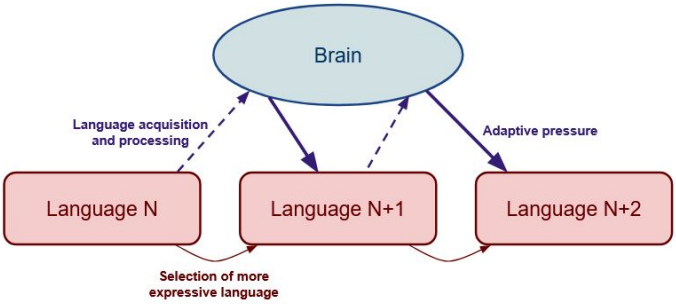


Figure 6: Figure adapted from [27]: Iterative adaptation of language to the brain structure and biases for optimized learning and processing.

speakers, capable of understanding each other. As a result, the authors argue that the properties of the human brain and cognition, as well as learning biases, are deeply embedded in language, so it can be easily assimilated and used (Figure 6). This would explain why natural languages are easier to assimilate than artificial ones [25]. In [58], Gibson *et al.* show that language evolved to be efficient from the perspective of information theory. This theory determines the efficiency of encoding a message as a function of the effort required by both the sender and the receiver to use it and transmit the information without loss, even when the physical transmission channel is subject to noise. One evidence of this is related to word size. In information theory, the more frequent a word, the shorter its encoding must be, so that the total size of the message is reduced (and therefore the communication effort). It has been verified that in the majority of languages, the most frequent words also tend to be the shortest [18]. To be even more precise, word size is strongly correlated with its predictability, which, unlike frequency (which is the same throughout the language), depends on the speaker and the context.

3.2.3 Nature of cognitive constraints

In their paper [27] Christiansen *et al.* describe the cognitive constraints on language that are likely to be also responsible for its evolution.

Thoughts are expressed through language and precedes any external expression, oral or written [26]. As mentioned above in Section 3.2.1, language is not merely a channel for communicating thought, but a

structural component in conceptualisation of thoughts. Language must therefore be optimised to develop, articulate, and express the thoughts of the human brain.

Perceptual-motor constraints are constraints on the physical transmission and reception of the signal. They are linked to the motor (vocal apparatus) [92] and sensitive (hearing) specificities of the human body.

Learning. As mentioned in Section 3.2.2, a language is considered fit from an evolutionary point of view if it can be learned efficiently. Christiansen *et al.* draw a parallel between learning a language and the brain's ability to learn from a complex sequence of stimuli, such as a discrete sequence of sounds for music or gestures for driving [28]. This learning process is called *"sequential learning"* in cognitive psychology. In both cases, the aim is to understand one element of the sequence in the light of those that precede and follow it [121, 123, 130] (in the case of language, the link between words, which are sometimes far apart), to extract meaning on the fly, and to break down the sequence into logical units (sentences in the case of language [33]). This particular ability of the brain may play a selection role (from an evolutionary perspective), leading to the emergence of languages that are increasingly adapted to this way of recording information. In addition, it appears that language learning and sequential learning make intensive use of the same area called Broca's area [133], and that there is a positive correlation between impaired sequential learning ability and impaired language ability [77, 169].

These clues suggest that language is shaped by the brain, and has evolved to be the most easily assimilated and usable for the human mind. Its structure is therefore profoundly influenced by the structure and limits of the brain. For this reason, in this thesis we studied the marks of cognitive limits that are made visible through language. We have therefore chosen to focus on the way in which individuals organise and use their vocabulary, in order to highlight common limitations in the ability to use a varied lexicon (Chapter 4) and topics (Chapter 5) in a stable manner over time (Chapter 6). We built on the work already done on limits in social relations and adapted the ego network model to the domain of language.

4 Structural invariants of ego networks of words

4.1 Introduction

As explained in Section 3.2, language is deeply tied to the way the human brain works: it powers internal information flows and suits its learning processes. As a result of this mutual connection, we can expect to find evidence of the structural limits of human cognition in language production. Specifically, we aim at investigating, through a data-driven approach, whether a regular structure can be found in the way people use words, as a "symptom" of cognitive constraints in this mental process. We argue that the usage of words might present properties similar to other mental processes which are known to be driven by cognitive constraints, specifically the way in which humans allocate cognitive capacity to maintaining social relationships, as seen in Section 3.1. In this section, we introduced the model called "ego network" that was used by Arnaboldi *et al.* [8] to uncover regularities in the way we organize our social life. The ego network arranges the relationships between an ego and its alters into concentric circles based on social proximity. Structural invariants were found in the number of circles as well as in their size, meaning that the human capacity to maintain social contacts is limited, and this limit is quantitatively similar for everyone [8, 44]. In this chapter we design an analogous model, the "ego network of words", to study how an individual (the ego) organizes the words he or she uses.

In trying to find out which cognitive constraints affect the production of language, our intuition is that we need to use records of spontaneous language. Indeed, it has been shown that the faster a task has to be completed, the greater the cognitive capacity required and the greater the risk of errors being made [126]. We argue that Twitter/X is a platform that facilitates a spontaneous writing style, much more so than newspaper articles (just to mention another textual formats readily available online). We collected a diverse dataset of tweets for our analysis, including tweets from regular Twitter/X users and professional writers (Section 4.2). Then, leveraging a methodology similar to the one used to uncover social constraints, we study the structural properties of language production on Twitter/X as a function of the individual word usage frequency, and we provide evidences for a set of cognitive constraints that naturally determine the way we communicate (Section 4.3). Specifically, our main findings 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) and of the specific time window considered.

- 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 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 clusters 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 layer consistently account for approximately 60% and 30% of the used words, irrespective of the number of clusters of the user.

4.2 Datasets

4.2.1 Data collection

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

Journalists Extracted from a Twitter/X list containing New York Times journalists[1], 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/X list created by Jennifer Frazer[2], 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 issued a tweet or a retweet in English with the hashtag

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

#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. As the 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 [35], which is based not only on language-agnostic features like the number of followers or the tweeting frequency, but also on linguistic features such as grammatical tags, or the number of words in a tweet [174]. The algorithm detects 29% of bot accounts, such that this dataset is composed of 5183 users.

Random users #2 This group has been collected by sampling among the accounts which issued 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. 23% of the accounts are detected as bot, such that this group contains 2733 accounts.

These groups are chosen to cover different types of users: the first two contain accounts that use language professionally (journalists and science writers), the other two contain regular users, which are expected to be more colloquial and less controlled in their language use. 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). In our analysis, we only consider active Twitter/X accounts, which we define as an account not abandoned by its user and that tweets regularly. Further details on this preprocessing step are provided in Appendix A.1.

4.2.2 Extracting user timelines with the same observation period

The observed timeline size is only constrained by the number of tweets (limited by API), thus the observation period 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. This raises the following problem: as 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 discussed in Section 6.3.1 (specifically, we cannot observe frequencies lower than the inverse of the observation period). In order to guarantee a fair comparison

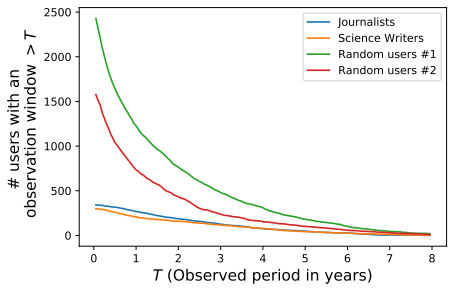


Figure 7: Number of selected timelines depending on the observation window.

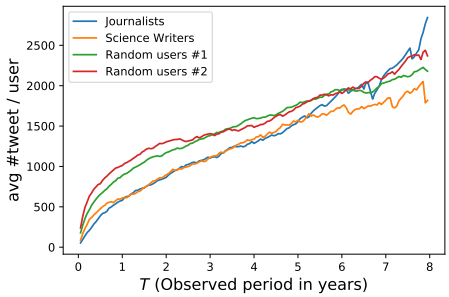


Figure 8: Average number of tweets depending on the observation window.

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 therefore reduces the number of profiles we can keep (see Figure 7): for a T larger than 2 years, that number is divided by two, 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 (Figure 8). 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 2 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). We, therefore, decided to carry on the analysis with $T \in \{1 \text{ year}, 2 \text{ years}, 3 \text{ years}\}$. Please note that random users have a higher frequency of tweeting than others. This difference tends to smooth out when the observation period is longer (Table 2). This can be explained by the fact that the timelines with the highest tweet frequency are excluded in that case because their observation period is too small.

4.2.3 Word extraction

Since the analysis has a focus on words and their frequency of use, we take advantage of NLP techniques for extracting them. As first step, all the

Datasets	Number of users				Avg # of tweets / user		
	1 year	2 years	3 years	4 years	1 year	2 years	3 years
NYT Journalists	268	187	125	75	579.71	865.02	1104.58
Science Writers	208	159	117	77	609.08	897.29	1112.63
Random Users #1	1227	765	481	311	897.29	1179.98	1403.50
Random Users #2	734	431	237	153	1057.41	1315.71	1404.60

Table 2: Number of users and tweeting frequency at different observation windows.

syntactic marks that are specific to communication in online social networks (mentions with @, hashtags with #, links, emojis) are discarded (see Table 12 in Appendix A.3 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. This type of words is called a functional word as opposed (in linguistics) 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 [38], and probably different neurological organizations [54]. We are more interested in lexical words because their frequency in written production depends on the author's intentions, as opposed to functional words frequencies that depend on the language characteristics[3]. Moreover, lexical words represent the biggest part of the vocabulary. Functional words are generally called stop-words in the NLP domain and many libraries provide tools to filter them out.

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 of overstemming and understemming. For this reason, we choose to perform lemmatization. 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. Table 13 in

[3]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 require a significant cognitive effort is arguable, hence in this work we opted for their removal.

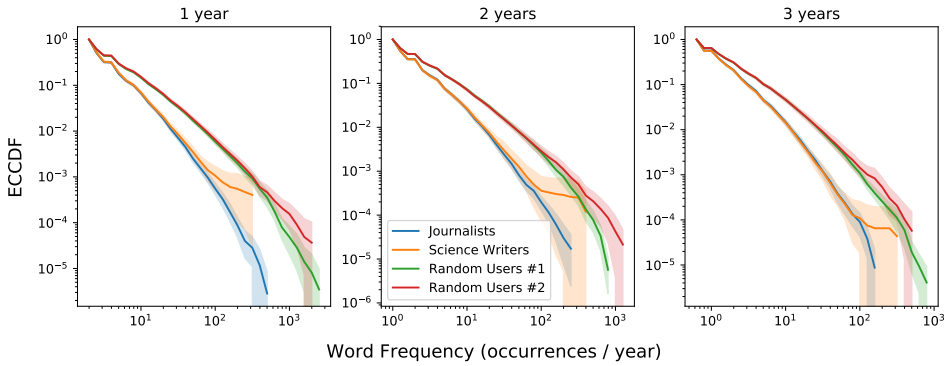


Figure 9: Aggregate visualization of user-wise empirical CCDF of word usage frequencies, in log-log scale. The solid line corresponds to the average and the shaded area to the 95% confidence intervals.

Appendix A.3 contains examples of the entire preprocessing part.

4.3 From word usage to cognitive constraints

4.3.1 Preliminaries

Let us focus on a tagged user j. When studying the social cognitive constraints, the contact frequency between two people was taken as 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 $\frac{n_{ij}}{t_i}$, where n_{ij} denotes the number of occurrences of word i in user j's timeline, and t_i denotes the observation window of j's account in years.

Figure 9 shows the frequency distribution for the different categories of users (regular users vs professional writers) and for the different observation windows (1, 2, 3 years). We can make two observations. First, the distributions exhibit a heavy-tailed behaviour. Second, the distributions are very similar two by two: specialized users (journalists and science writers) who fulfill a particular role of information in the social network, and randoms users who are samples of more regular users. The first group seems to use more low-frequency words, while the second group uses a larger proportion of high-frequency words. Based on the second observation, we can compare the datasets based on two criteria: verbosity, which counts the total number of words per tweet, and *lexical richness*, which counts the number of distinct words per tweet (Figures 10-11). Despite a lower verbosity (Figure 10), the

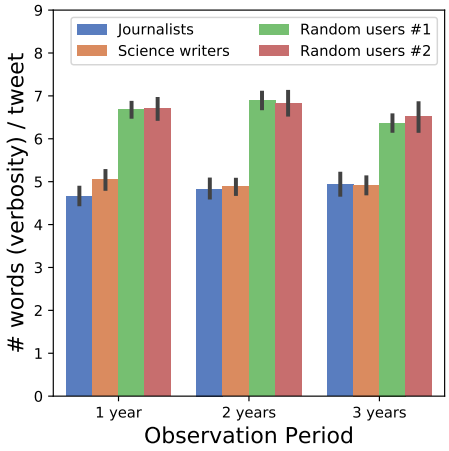


Figure 10: Average verbosity, with 95% confidence intervals.

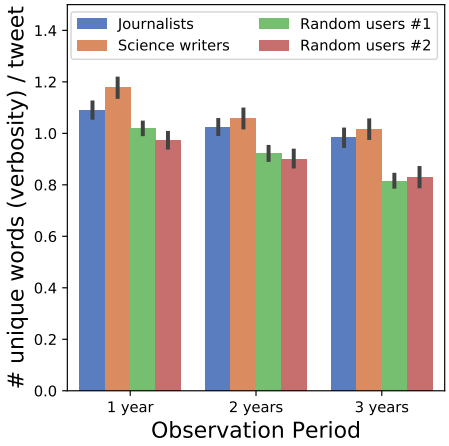


Figure 11: Average lexical richness, with 95% confidence intervals.

vocabulary of specialized users seems richer than those of random users (Figure 11). This is intuitive, but not necessarily expected. Professional users certainly have a higher diversity in the use of words. However, this manifests also in Twitter, i.e., in an environment where they do not write necessarily for professional reasons, but where they are (supposedly) writing in a more immediate and informal way.

4.3.2 Many words, just a few groups

Using the frequencies described in the previous section, we now investigate whether the words of a user can be grouped into homogeneous classes, and whether different users feature a similar number of classes or not. To this aim, for each user, we leverage a clustering algorithm to group words with a similar frequency. The selected algorithm is Mean Shift [55], because as opposed to Jenks [83] or KMeans [107], it is able to find the optimal number of clusters without fixed parameters. The original Mean Shift algorithm has nevertheless a drawback: it is only able to find the estimated density peaks with a fixed bandwidth kernel. The bandwidth is estimated based on the distribution of the pairwise distances between all the frequency values. However, in our case the distance between frequencies is not homogeneous: most of them are concentrated in the lowest values, close to each other. Hence, the selected bandwidth is fitted for estimating the density in that area, but not in the tail of the distribution. For that reason, a log-transformation is applied to

the frequency values prior to the Mean Shift run: it still allows a fine mode detection in low-frequency part and compresses high values to allow detection of modes with a larger width. The use of a logarithmic scale is also used by psychological researchers to explain the impact of word frequency on their cognitive processing [17].

The histograms of the obtained optimal number of clusters are shown in Figure 12. It is interesting to note that, despite the heterogeneity of users (in terms of tweeting frequency, verbosity, and lexical richness), the distributions are always quite narrow, with peaks consistently between 5 and 7 clusters. The observation period seems to have a very limited effect on the resulting cluster structure. This means that, after one year, the different groups of words can be already identified reliably. In addition, this limited effect actually reinforces the idea of a natural grouping: when more words are added (longer observation period) the clusters become slightly fewer, not slightly more. Hence, new words tend to reinforce existing clusters. Thus, 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 chapter, hinting at the existence of structural invariants in cognitive processes, which we summarise below.

Cognitive constraint 1: 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 [131], 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 to or by the observation window. This regularity might also suggest that these groups of words correspond to linguistic functional groups, and we plan to investigate this as future work.

4.3.3 Exploring the group sizes

We now study the size of the clusters identified in the previous section. For the sake of statistical reliability, we only consider those users whose optimal number of clusters (as identified by Mean Shift) corresponds to the most popular number of clusters (red bars) in Figure 12. This allows us to have a sufficient number of samples in each class. 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. Following the convention of the Dunbar's model discussed in

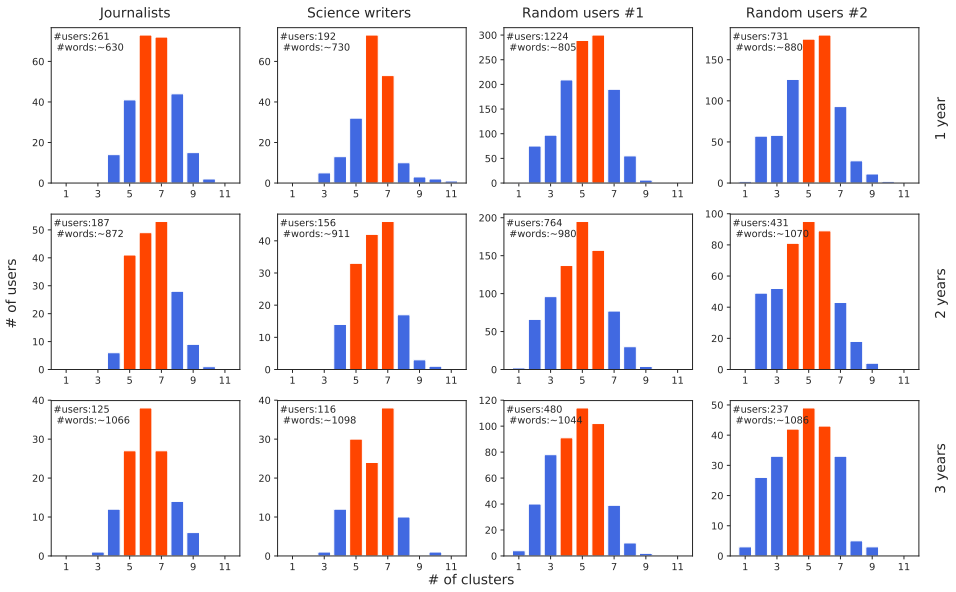


Figure 12: Number of clusters obtained applying Mean Shift to log-transformed frequencies. The most frequent number of clusters are highlighted in red.

Section 3.1.3, these clusters can be mapped into concentric layers (or circles), which provide a cumulative view of word usage. Specifically, layer 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 *ego network of words*.

Figure 13 shows the average layer sizes for every dataset and different observation periods. As expected, for a given number of clusters, the layer size increases as we expand the observation period, because more words are brought in. For a given number of clusters we also observe a striking regularity across the datasets, with confidence intervals overlapping in practically all settings. Typically, the layer sizes are slightly higher for journalists and science writers ($T = 2$ years and $T = 3$ years). The main reason is that their lexicon is generally richer than those of regular users (as discussed in Section 6.3.1) and this is reflected in their layer size.

Another typical metric that is analysed in the context of social cognitive constraints is the scaling ratio between layers, which, as discussed earlier, corresponds to the ratio between the size of consecutive layers. The scaling ratio is an important measure of regularity, as it captures a relative pattern

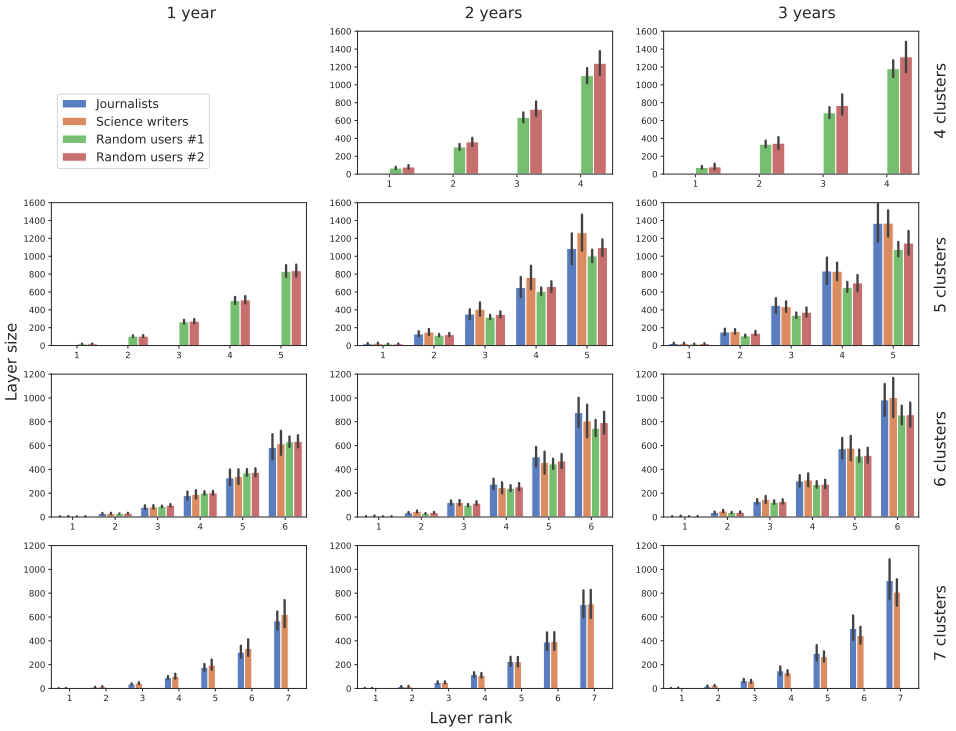


Figure 13: Average layer size (with 95% confidence intervals) for the various datasets, different number of clusters (rows), and different observation periods (columns). For reasons of clarity, only the ego networks with the most frequent numbers of layers (in red in Figure 12), for a given dataset and observation period, are plotted. This explains why the first cell is empty.

across layers, beyond the absolute values of their size. Figure 14 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. 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 word 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 (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 with T=1 year (but analogous results are obtained for the other categories and observation windows) and in Figure 15 we plot the linear regression for all the user categories with $T = 1$ year. 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 (Figure 15). Moreover, it is quite interesting to observe that the second-last and third-last layer 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 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 Figure 14, 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). We leave as future work the characterization of this special layer and we summarise below the main results of the section.

Cognitive constraint 2: Structural invariants in terms of layer sizes and scaling ratio are observed also in the language domain. 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.

Cognitive constraint 3: Users with more clusters organise differently their innermost clusters, 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.

4.4 Conclusion

We investigated, through a data-driven approach, whether a regular structure can be found in the way people use words, as a symptom of cognitive constraints in their mental process. This is motivated by the fact that other mental processes are known to be driven by cognitive constraints, such

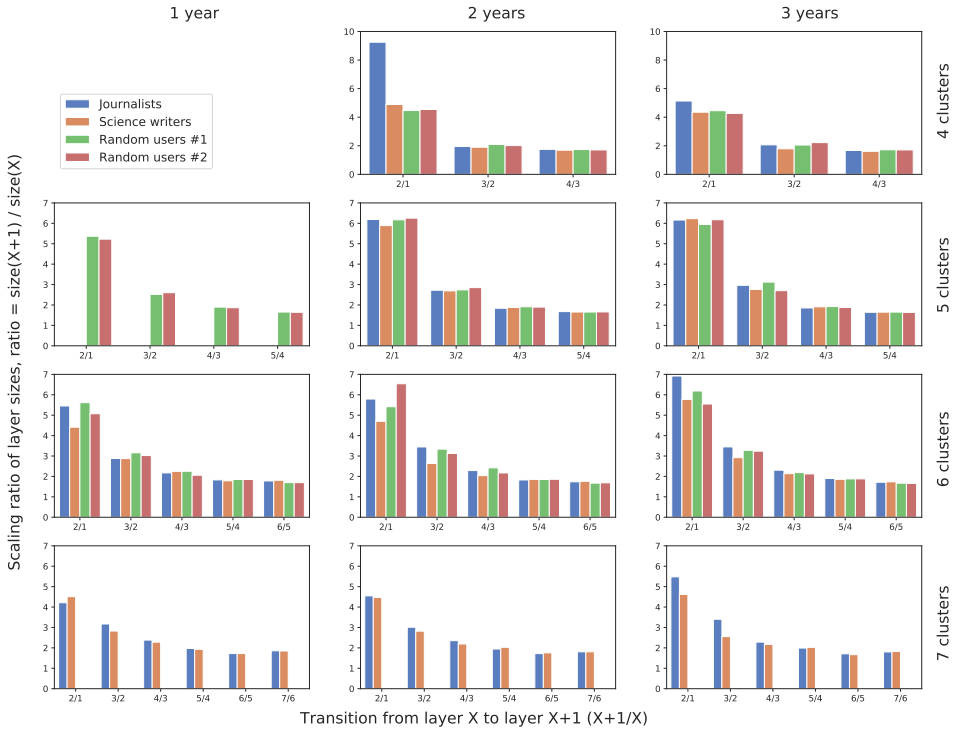


Figure 14: Scaling ratio for the various datasets, different number of clusters (rows), and different observation periods (columns).

Opt. # of clusters	Cluster Rank						
	1	2	3	4	5	6	7
4 clusters	0.04	0.33	0.61	1.00			
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

Table 3: Linear coefficients obtained for the journalists dataset with $T = 1$ year.

as the way how humans allocate cognitive capacity to social relationships. To this aim, we collected a diverse dataset from Twitter/X (identified as one of the major sources of informal and spontaneous language online), including tweets from regular Twitter/X users and from professional writers. Then, leveraging a methodology similar to the one used to uncover social constraints, we have analysed the structural properties of language production on Twitter/X, uncovering regularities that constitute preliminary evidence of the aforementioned cognitive constraints. Specifically, we have found that, similarly to the social case, a concentric layered structure (ego network of words) very well captures how an individual organizes their cognitive effort in language production. Words can be grouped typically in between 5 and 7 layers, regardless of the specific class of users. We also observe a structural invariant in the size of the layers, which grow approximately 2-3 times when moving from a layer to the next one. Another structural invariant emerges for the external layers, which, regardless of the number of clusters of the user, consistently account for approximately 60% and 30% of the used words.

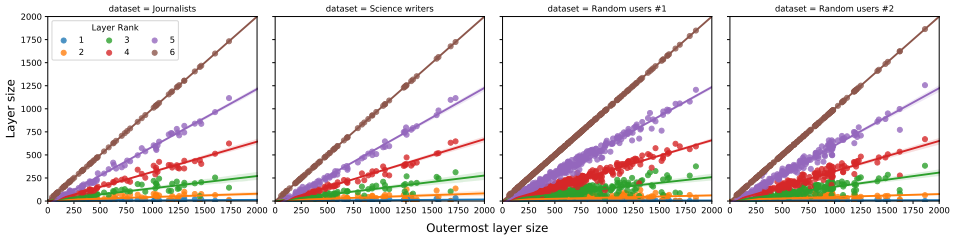


Figure 15: Linear regression between the total number of unique words used by each user (corresponding to the size of the outermost layer) and the individual layer sizes.

5 Semantic invariants of ego networks of words

5.1 Introduction

In the previous chapter, we treated words as units of language through the prism of frequency. However, we know that the way the brain processes words depends not only on their frequency, but also on their semantic value. For example, in experimental conditions [50], a word is recognised more quickly when it is preceded by a semantically related word (such as "whale" and "dolphin"). We thus complement in this chapter the structural analysis with a semantic study. We rely on the very same datasets extracted from Twitter/X in the previous chapter. We made this choice because it allows us to study a wide range of people with different profiles and because of its favorability for spontaneous and natural reactions. To this end, we classified the words into topics and then studied the importance of these topics in the different rings of ego networks obtained in the previous chapter. This extensive semantic analysis of ego networks of words also highlights ways to characterise specificities of individuals as they emerge from their use of words.
The key findings of this chapter are the following.

- 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.

5.2 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 5.2.1, we motivate our selection of the BERTopic framework for topic extraction. Then, in Section 5.2.2, 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 5.2.3 we build the *semantic profile* of the ego network ring, as the topic distribution of the words in that ring.

5.2.1 Preliminaries

To calculate a semantic profile, we choose to consider the meaning of each word in its context rather than using a semantic dictionary [146] (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 [53], which is a semantic web database extracted from a news portal that can be used for news recommendation purposes [84], or the BBC ontologies extracted from the BBC corpus of news, which allows politically-oriented topic mining [1]. 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 [111], we prefer having the maximum freedom in the topic identification process by using a transformers-based model such as BERT [36] which is the current state of the art in text embedding and then using an unsupervised method to detect topics.

5.2.2 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 [69], then all word occurrences that constitute a tweet are assigned the same topic as the tweet itself (Figure 16).

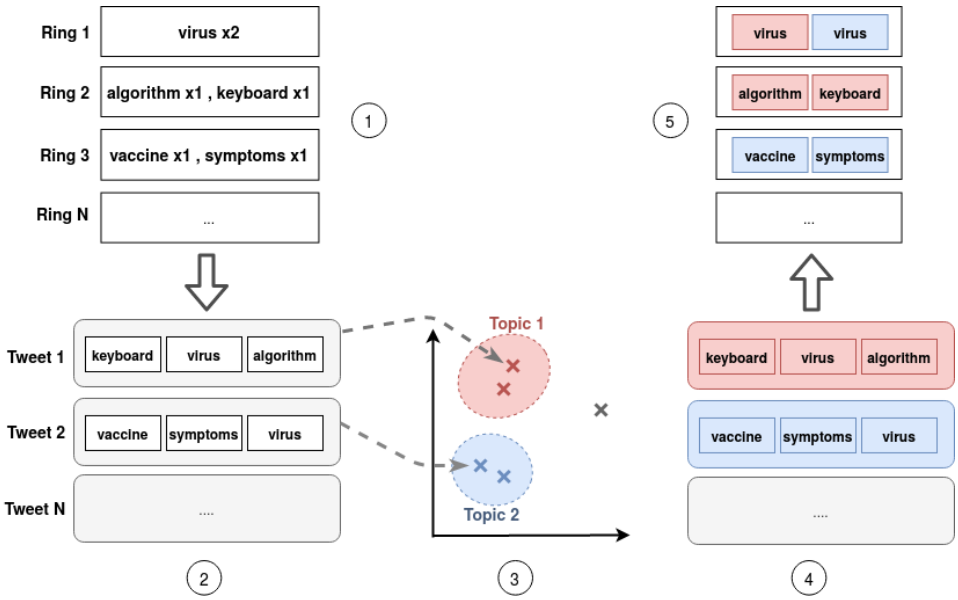


Figure 16: **Obtaining the semantic profile of the rings of an ego network.** (1) The ego network's rings organize a user's vocabulary based on the frequencies of the words. (2) For a given word, its occurrences in the user timeline are coming most likely from different tweets. (3) The tweets are classified 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.

For the current analysis, we chose to focus only on ego networks with six rings (with an observation period of one year) the case covering the most users. As described in the following, the BERTopic framework uses sequentially BERT [36] for tweet embedding, UMAP [110] for dimension reduction, and HDBSCAN [109] for clustering those tweet embeddings in a low-dimensional subspace.

Tweet embedding with BERT.

BERT [36], 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 [187] 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.

Dimensionality reduction with UMAP.

In order to mitigate the curse of dimensionality (to which clustering algorithm based on k-nearest neighbors are particularly sensible [139]), 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 embedding space down to five dimensions as recommended in the BERTopic framework [69]. UMAP, like the T-SNE [173] algorithm, is able to capture latent non-linear dimensions but in a more scalable way.

HDBSCAN for clustering topics.

HDBSCAN [109] is also able to find non-linear cluster structures from the density, as well as outliers, like DBSCAN (Figure 17). 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 4 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 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

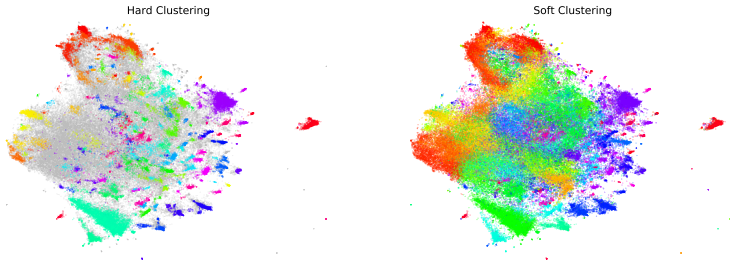


Figure 17: **2D visualization of the HDBSCAN results on the Journalists dataset with both hard and soft clustering.** 265 clusters are found (they are the same in both cases). In the first case, each point is classified as either belonging to a single cluster (colored points) or as an outlier (grey point), 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).

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 Appendix A we discuss in detail why hard clustering is better suited for our analysis.

Table 4: **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%

Reduction of the number of topics

As shown in Table 4, 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 Figure 18. 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 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.

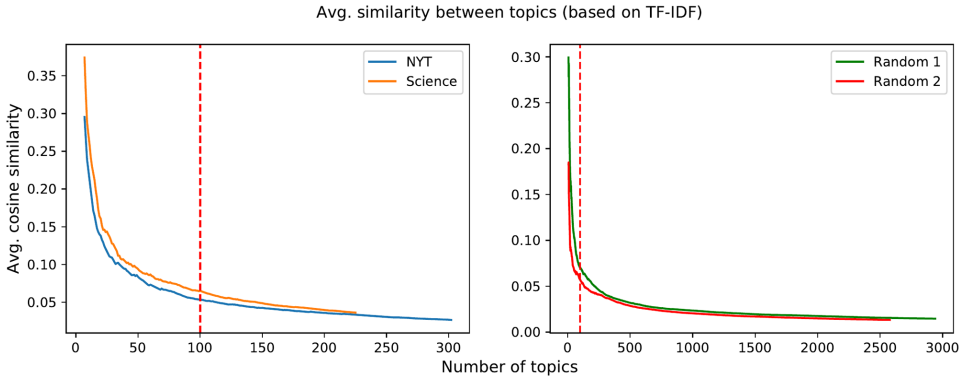


Figure 18: **Number of topics vs. average topic similarity.** The threshold of one hundred topics is marked with the dashed red line. This threshold is situated at the end of the bend for specialized datasets, and in the middle of the bend for both random datasets.

5.2.3 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 Ω_{e,r_i} be the set of word occurrences contained in ring r_i of the ego network e, and $m(w)$ the tweet the word occurrence w belongs to. The probability $P_{e,r_i}(c)$ of observing topic c in ring r_i of ego network e is defined as follows:

$$P_{e,r_i}(c) = \frac{\sum_{w \in \Omega_{e,r_i}} P_{m(w)}(c)}{\sum_{c \in \mathcal{C}} \sum_{w \in \Omega_{e,r_i}} P_{m(w)}(c)} \tag{1}$$

where $\sum_{c \in \mathcal{C}} P_{e,r_i}(c) = 1$. More in general, we denote with P_{e,r_i} the semantic profile of ring r_i in ego network e (depicted in Figure 19). For this reason, we will also refer to $P_{e,r_i}(c)$ as the share of c in the semantic profile P_{e,r_i} of r_i This unique semantic profile will be the starting point for all subsequent analyses in this section. In Appendix A, we provide four tables (one for each dataset) that detail for every topic the most characteristic words and the average share in the rings.

Note: Two different semantic profiles can be built, depending on whether topics are assigned using hard vs soft clustering. In Appendix A 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 5.4 only the results obtained with hard clustering. In Appendix A we discuss soft versus hard clustering in detail and motivate why hard clustering is better suited for our analysis.

5.3 Metrics for the analysis of semantic profiles

After following the steps described in Section 5.2, we end up with a semantic profile for each ring of an ego network. In the following we discuss (i) how to

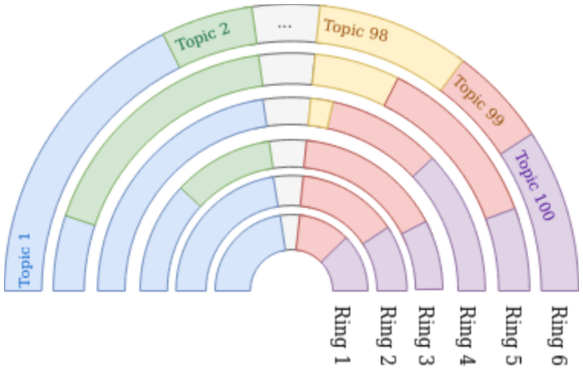


Figure 19: **Semantic profile illustration**. Each ring is associated with a topic distribution.

characterise individual semantic profiles (Section 5.3.1), (ii) how to compare semantic profiles (Section 5.3.2), and (iii) how to leverage semantic profiles to investigate the role of the most important topics (Section 5.3.3).

5.3.1 Characterization of the semantic profile

Let us consider a ring r_i 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_i touch upon. Formally, the number of topics associated with a given ring can be calculated as follows:

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

where we denoted with $P_{e,r_i}(c)$ the probability of a observing topic c in the semantic profile P_{e,r_i} of ring r_i, and $\mathbb{1}$ is the indicator function. Note, though, that $\mathcal{N}_{e,r_i}$ may offer only a partial perspective. In fact, rings have very different sizes (as discussed in Chapter 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 $\mathcal{N}_{e,r_i}$ with its normalised version:

$$\mathcal{N}'_{e,r_i} = \frac{\mathcal{N}_{e,r_i}}{|\Omega_{e,r_i}|}, \tag{3}$$

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

$\mathcal{N}_{e,r_i}$ and $\mathcal{N}'_{e,r_i}$ 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_{e,r_i}. Recalling that P_{e,r_i} 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 $P_{e,r_i}(c)$), 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_i} the entropy of the ring r_i in ego e, its definition is as follows:

$$H_{e,r_i} = -\sum_{c \in \mathcal{C}} P_{e,r_i}(c) \times \log P_{e,r_i}(c). \tag{4}$$

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

In Section 5.4, the average of $\mathcal{N}_{e,r_i}$, $\mathcal{N}'_{e,r_i}$, and H_{e,r_i} across all ego networks will be presented, i.e. $\mathcal{N}_{r_i} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathcal{N}_{e,r_i}$ (analogously for the others).

5.3.2 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 [100], 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 [125]. 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_{e,r_i} and P_{e,r_j} of rings i and j for ego network e can be computed as follows:

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

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

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

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

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

5.3.3 Capturing important topics and their cross-rings effects

Given a semantic profile P_{e,r_i}, 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 $P_{e,r_i}(c)$ where $c \in \mathcal{C}$. To this aim, we use the Jenks algorithm [82] 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 [142] 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_{e,r_i} (where e is the ego network and i is the ring number), and the latter is the set of non-important topics as L_{e,r_i}.

Once we have obtained U_{e,r_i} and L_{e,r_i}, 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 $K^{r_y}_{TOP(r_x)}$ 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.

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

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

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

Basically, $S^{r_y}_{TOP(r_x)}$ measures the average share of *each* r_x's primary topics in another ring of the same ego network. Similarly, we can compute

$S^{r_y}_{BOTTOM(r_x)}$ by replacing U_{e,r_x} with L_{e,r_x} 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:

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

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:

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

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 $\sigma^{r_y}_{TOP(r_x,r_y)}$.

5.4 Results

In this section, we study the semantic profiles in the ego networks of the Twitter/X users in our four datasets.

5.4.1 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 ($\mathcal{N}_{e,r_i}$ and $\mathcal{N}'_{e,r_i}$, 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. 20 (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. 20 (b)): as seen previously in Section 5.2.2, 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 ($\mathcal{N}'_{r_i}$), the maximum of the normalised topic count (Fig. 20 (c)) is observed in the first ring. Thus, *ring #1 stands out as the ring that generates proportionally more topics than the other rings.*

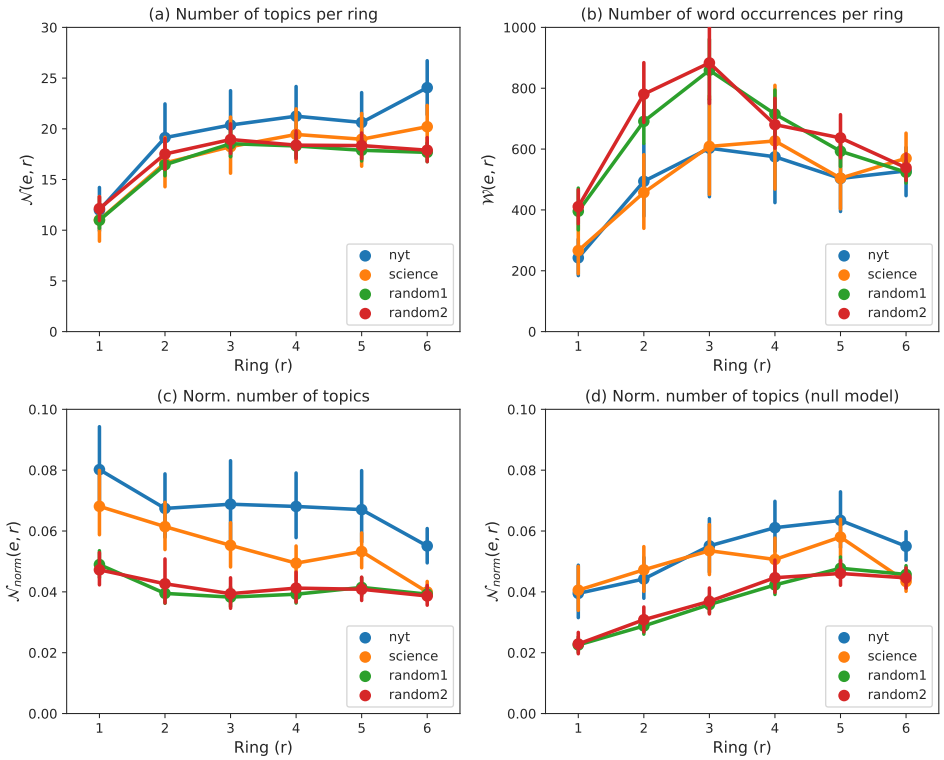


Figure 20: Average number of topics (a), number of word occurrences (b), and normalised number of topics (c) in each ring of the ego network. For "null" ego networks, we report only the normalised number of topics (d).

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. 20 (d), we show $\mathcal{N}'_{r_i}$ for the null model of ego networks. Since the maximum of $\mathcal{N}'_{r_i}$ is obtained at a different ring r_i 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.

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_{e,w_u} be the number of occurrences of the word w_u in ego e, such that the number of word occurrences in a ring r_i of a given ego e is defined as:

$$O_{e,r_i} = \sum_{w_u \in \mathcal{W}_{e,r_i}} o_{e,w_u} = |\Omega_{e,r_i}|, \tag{13}$$

$\mathcal{W}_{e,r_i}$ being the set of unique words in ring r_i. For each ego network, all the words are shuffled (i.e., a new $\mathcal{W}'$ 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:

$$\begin{cases} |\mathcal{W}'_{e,r_i}| = |\mathcal{W}_{e,r_i}| \\ O'_{e,r_i} = O_{e,r_i}. \end{cases} \tag{14}$$

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:

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

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 $P_{m(w_c)}(c) = 1$. 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_i is the weighted average of the topic distribution P_{e,w_u} of the unique words $w_u \in \mathcal{W}'_{e,r_i}$ that compose that ring after shuffling

$$P_{e,r_i}(c) = \frac{\sum_{w_u \in \mathcal{W}'_{e,r_i}} o'_{e,w_u} \times P_{e,w_u}(c)}{\sum_{w_u \in \mathcal{W}'_{e,r_i}} o'_{e,w_u}}. \tag{17}$$

The full process is summarized with a toy example in Fig. 21.

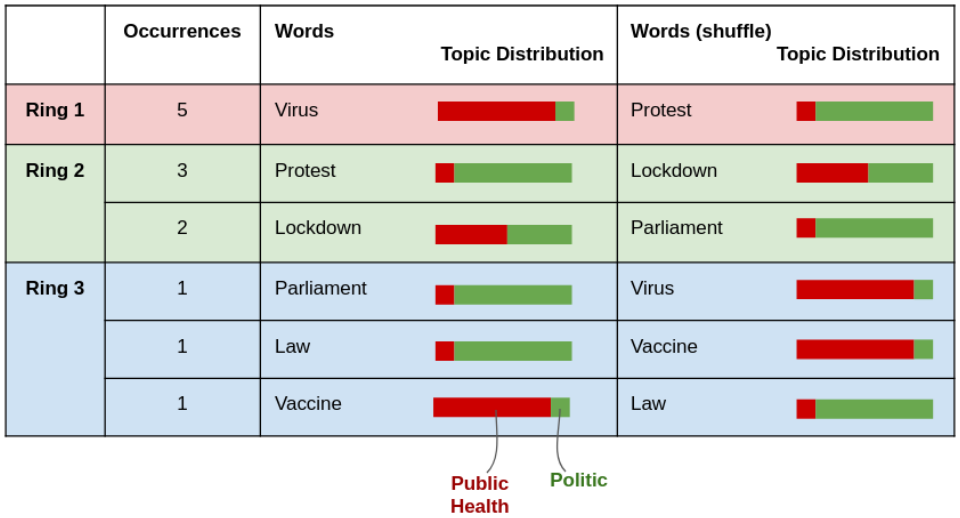


Figure 21: **Null model example**. The ring sizes and word occurrences are kept, the words are shuffled. In this toy example: $O_{e,r_2} = 3 + 2$, $o_{e,virus} = 5$, $o'_{e,virus} = 1$.

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 5.3.1. 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. 22 (left) shows different levels of entropy depending on the rings: H_{r_i} 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. 22 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.

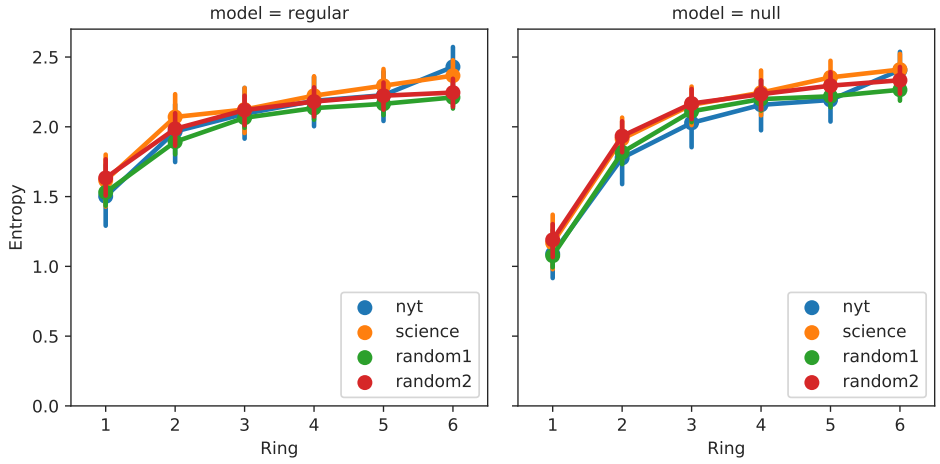


Figure 22: **Entropy of the semantic profiles per ring.** Real-life ego networks (left) vs null model ego networks (right).

We now carry out a pairwise comparison of the semantic profiles of rings, using the JS distance described in Section 5.3.2. we plot the, in Fig. 23. 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.*
- 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/X 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. 20 (c)), but the distribution in ring #1 is the furthest away from the others* (Fig. 23). 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.

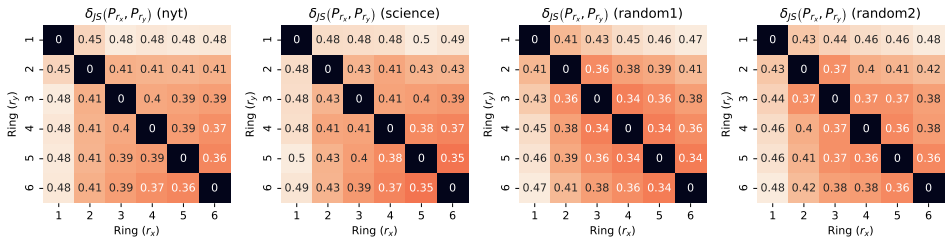


Figure 23: **Jensen-Shannon distance.** Average JS distance between the rings.

Take home message for Section 5.4.1: 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.

5.4.2 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 5.4.3.

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 5.3.3), varying r_y from innermost to outermost layer. Fig. 24 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.

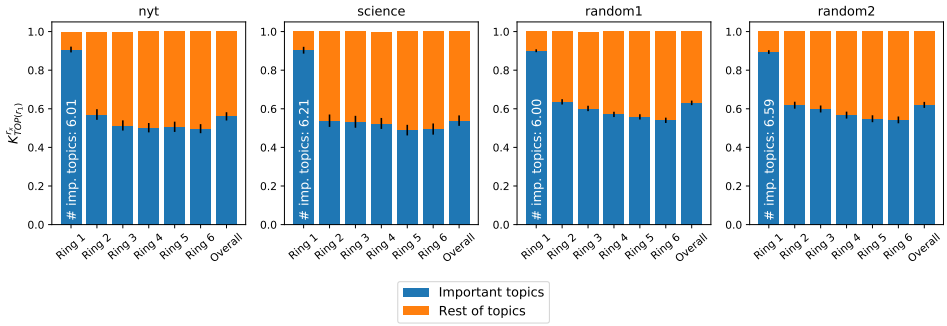


Figure 24: **Average strength of ring #1's important topics in the semantic profile of each ring and of the whole ego network.** Each bar stands for the semantic profile of each ring (and overall ego network, in the last bar), where the blue part represents the share covered by the most important topics of ring #1 (their average number $|U_{r_1}|$ is written in white).

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 $K^{r_y}_{TOP(e)}$. Figure 25 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. 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/X user is very likely to belong to U_{e,r_1}.

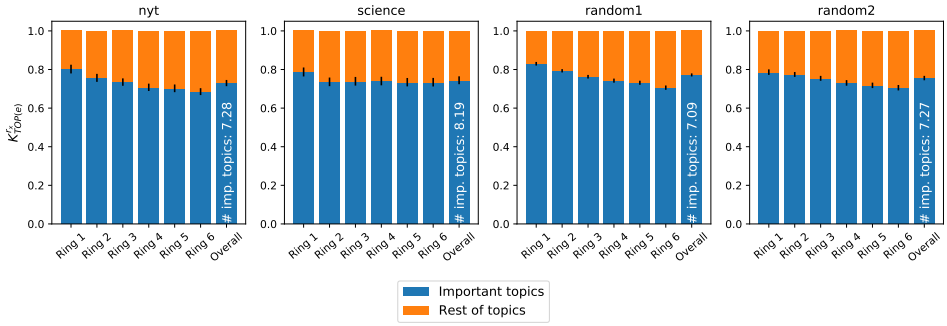


Figure 25: **Average strength of the ego network's important topics in the semantic profile of each ring.** The blue part of the stacked bar represents the share covered by the important topics in U_e. The average number of topics $|U_e|$ is specified in white.

Take home message for Section 5.4.2: Both results from Fig. 24 and 25 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 5.4.1), confirming the special role of this ring in the ego network of words.

5.4.3 Pulling power of primary topics

Let us now focus on the primary topics in a generic ring r_x (i.e., those in U_{e,r_x}). They can also appear in another ring r_y, and can be found in either U_{e,r_y} or L_{e,r_y}. 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 the notation of Section 5.3.3, 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 5. 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 5.4.2), 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 5.

From the left side of Table 5, 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 6. 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 $p - \text{value} < .001$. 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 5.4.3: 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

Table 5: **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	

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.

Table 6: **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	

5.4.4 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 [7], or to perform link prediction [171]), 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/X 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.

5.5 Conclusion

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.

6 Extracting "active" ego networks of words

6.1 Introduction

As introduced in the previous chapters, the ego network of words is a novel model that captures structural properties in language production linked to cognitive constraints. We already studied its layer-based structure (Chapter 4) and its semantic properties (Chapter 5). In this chapter, 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[4].

In Chapter 4, 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 section, 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 chapter 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 chapter 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 in this chapter is the definition of a methodology to extract the "active" part of the ego network (Section 6.3). In Section 6.4, we successfully test this methodology using two types of datasets: interview transcripts and tweets. MediaSum [186] is a dataset that includes thousands of verbatim transcripts of spoken interviews from an American public radio and private TV channel (Section 6.2.1). The Twitter/X datasets are extracted

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

from the same users as in Chapters 4 and 5, but we downloaded larger timelines, up to 10K tweets (Section 6.2.2). We also prove that the method that we use to extract the active ego network is robust to different amounts of input data (Section 6.4) and that the active size is stable over time The structural results (Section 6.4.2) of the ego networks produced in this way substantially confirm the layer ego network of word structure obtained in Chapter 4 but are robust to the size of the input data. The second contribution of the chapter is the validation of the ego network of words model on a dataset (MediaSum [186]) that is completely different in nature from the Twitter/X ones on which it had been applied previously. 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 chapter 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.

- 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 chapters (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.

6.2 Datasets

In this study, we will rely on two types of datasets. The first, MediaSum [186], compiles years of television and radio interview transcripts. The second is an extension of the Twitter/X datasets used in previous chapters, for which we have extended the number of messages collected per person to 10K.

6.2.1 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 26). 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 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 7). In our previous chapters, where we only 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

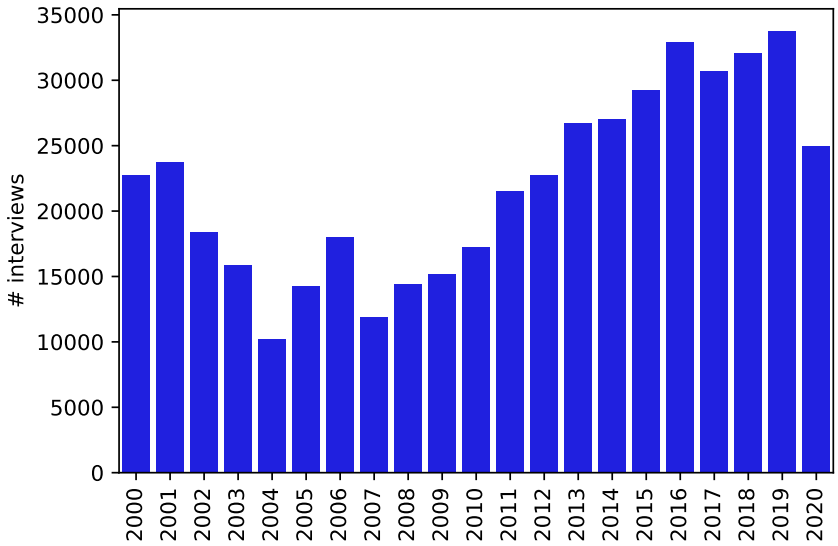


Figure 26: Number of interview transcripts per year in the MediaSum dataset.

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 27 and Figure 28, a few speakers have a very large number of word occurrences and unique words. Unsurprisingly, most 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.

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

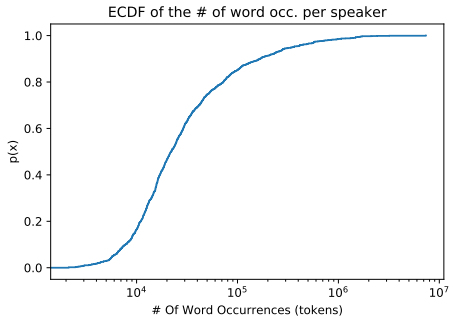


Figure 27: Word occurrences per speaker

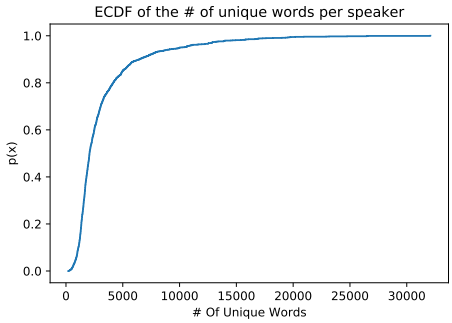


Figure 28: Unique words per speaker

6.2.2 Twitter/X

In previous chapters, we built ego networks of words based on Twitter/X timelines with up to 3.2K tweets (the download limitation of the legacy standard Twitter API) collected from four sets of users (Section 4.2.1): journalist working for the New York Times, science writers who tweet about science-related topics, and two distinct sets of random users. 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 29, this results in much longer timelines with respect to those analysed in previous chapters. These longer timelines are used to stress-test the ego network of words model. In the same fashion as in Chapters 4 and 5, and in Section 6.2.1, we only keep the timelines with at least 500 tweets. The figures related to the number of word occurrences and

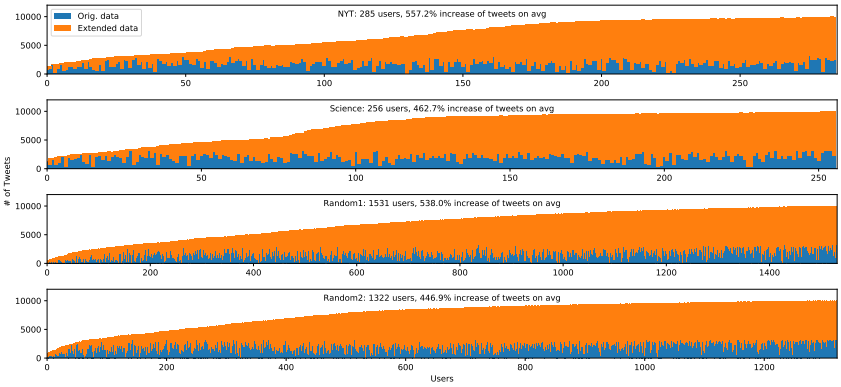


Figure 29: Collected Twitter/X timelines containing at least 500 tweets. Each bar corresponds to a timeline, where the blue part refers to the number of tweets in the original dataset, and the orange part refers to the number of newly collected tweets.

unique words are reported in Table 8. 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.

Dataset	# of users	Avg. word occurrences / user	Avg. unique words / user
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 8: Twitter/X datasets after removing users with less than 500 tweets

6.3 Methodology

6.3.1 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 9). 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 6.2.1, for each ego e we end up with a tuple (i.e., an ordered sequence) of tokens [116], 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 [116]. 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 chapter, 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 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.

6.3.2 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 already presented in Chapter 4.3. 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, ... , r_τ, 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:

$$\mathcal{W} = \bigcup_{i=1}^{\tau} \mathcal{W}_{r_i}. \tag{18}$$

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

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 9: Summary of notation used in the chapter.

$j \leq i$. The set of words assigned to layer l_i is denoted with $\mathcal{W}_{l_i}$, so:

$$\mathcal{W}_{l_i} = \bigcup_{j=1}^{i} \mathcal{W}_{r_j}. \tag{19}$$

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 [55] algorithm, which is able, in contrast to Jenks [83] or K-Means [107], to automatically optimize τ, the number of groups to be found. 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.

6.3.3 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 30. 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 chapters (usually between five and seven circles), despite exactly the same workflow being used. Note also that the Twitter/X datasets used here are *the same* as those in Chapters 4 and 5 and except for the timeline length considered (much larger, in this work). As we can observe in Figure 31, 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 [8]. The conventional cut-off point, as stated in [41], 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 6.3, in order to discard the words that do not take up cognitive capacity.

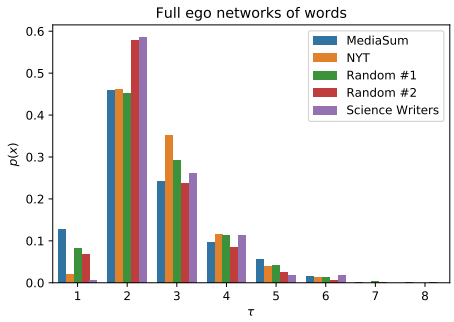


Figure 30: Distribution of the number of circles τ when considering all the words available in $\mathcal{W}$

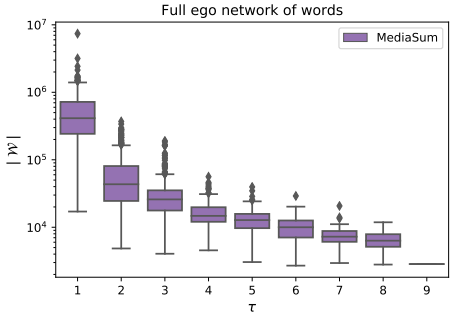


Figure 31: Number of circles τ vs full ego network size ($|\mathcal{W}|$) in the MediaSum dataset. The same trend is observed in the other datasets (plots omitted to optimize the space).

6.3.4 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 32, 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 6.3.1, $s : n \mapsto w^n$.

In Figure 32 and 33, we present two typical cases observed in our datasets. Figure 32 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 33 is representative of users who were not observed for a sufficient duration to reach the second phase described in Figure 32. 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 32.

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 32, 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, outlined in Section 6.3.5, 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 6.4. Note that sometimes the textual data for one ego is not large enough for the ego network to reach any cut-off point (Figure 33). 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.

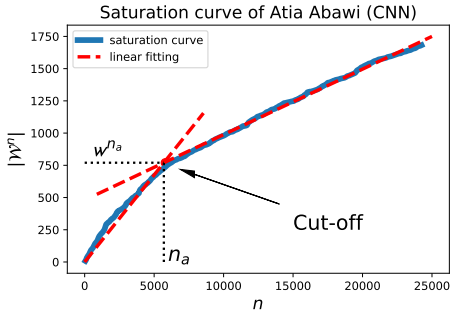


Figure 32: Non-linear saturation curve.

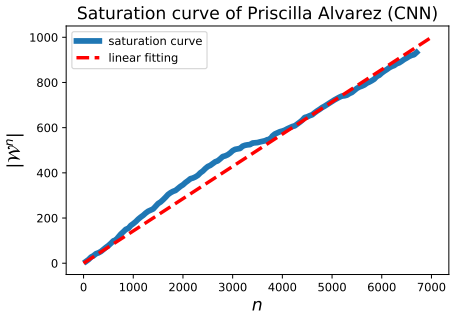


Figure 33: Linear saturation curve.

6.3.5 Methodology for identifying the cut-off point

We start with a high-level description of our methodology, illustrated in Figure 34. 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 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.

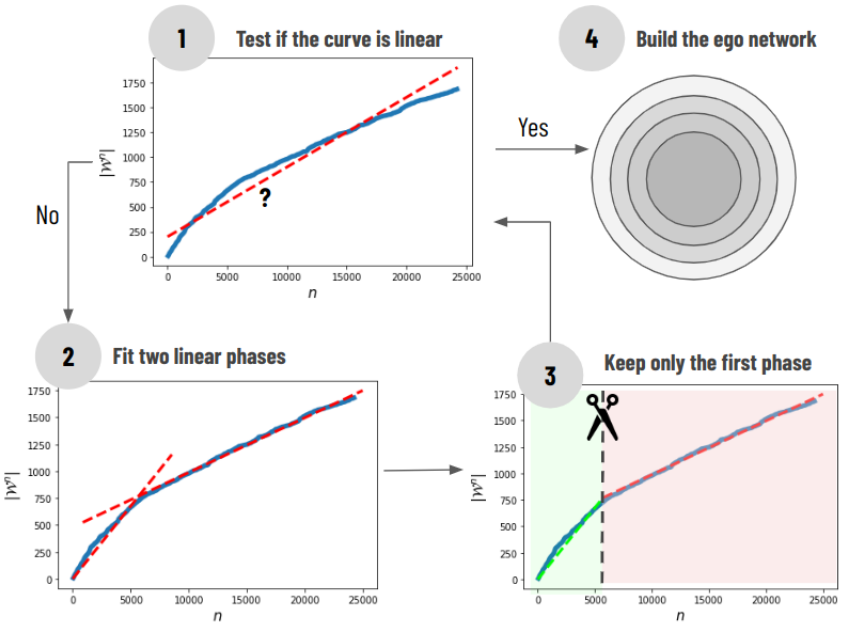


Figure 34: Steps for detecting the saturation point. 1) Linearity test. 2) If the curve is not linear, we find the best model fit with two linear parts. 3) Keep only the part of the curve which fits the first linear part of the model, then come back to 1).

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 [167] for its ability to operate a variable reduction on its input features. The features used by Lasso are the polynomial terms of 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 nonzero 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 35 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.

6.4 Results

The goal of this section is to fully validate the methodology proposed in Section 6.3. First, in Section 6.4.1 we show that the layered structure that was not present when considering the full ego network (Figure 30) emerges again when focusing on the active ego network, and we revisit its properties in Section 6.4.2. Then we evaluate the robustness of the methodology to a varying amount of input data (Section 6.4.3). Finally, we show that active

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

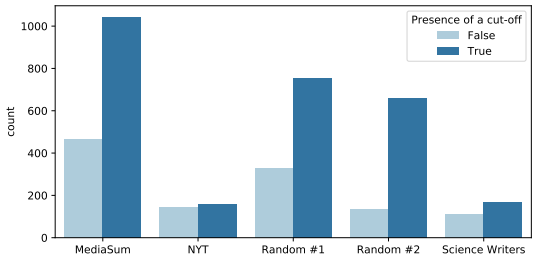


Figure 35: Amount of egos with and without a cut-off point.

ego networks are stable over time

6.4.1 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 30. By employing the methodology outlined in Section 6.3, we can now effectively isolate[5] the active component of the ego network and ascertain whether the layered structure reemerges. Figure 36 demonstrates that this is indeed the case. Comparing it with Figure 30, 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 6.3 is able to properly identify the active part.

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 10, 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 10 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.

[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.

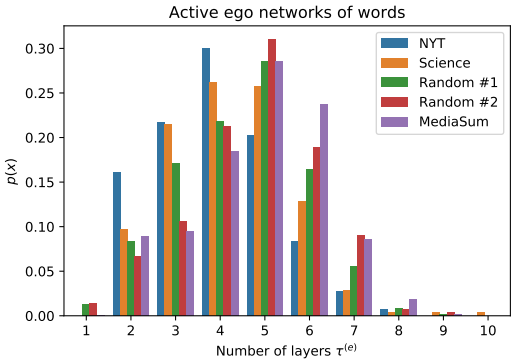


Figure 36: Distribution of the number of circles for the active ego networks of words.

When we consider the results from Figure 36 in conjunction with Table 10, 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.

6.4.2 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 6.2. 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 6.3. From now on, we simplify the notation w^{n_a} to w.

The analysis in Figure 36 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/X users #2. Interestingly, a similar optimal range of 4 to 5 circles was also observed in the social domain [44].

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 36, hence providing more samples

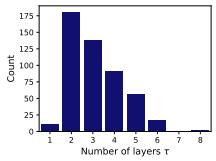
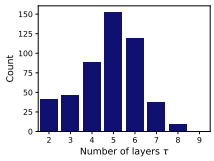
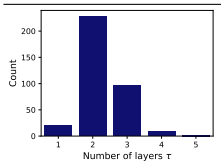
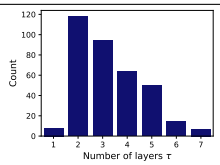
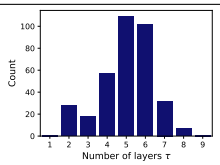
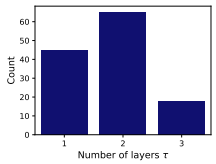
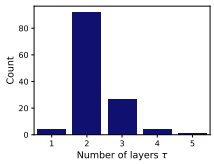
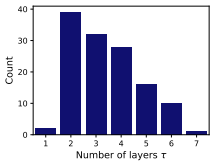
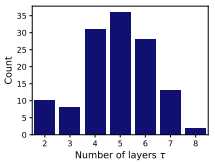
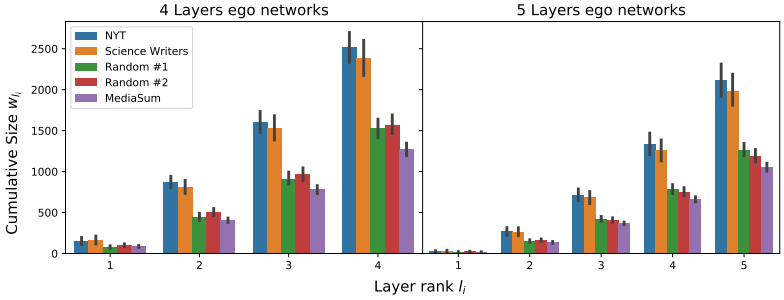
No cut	1st iteration	2nd iteration	3rd iteration
			
			
			

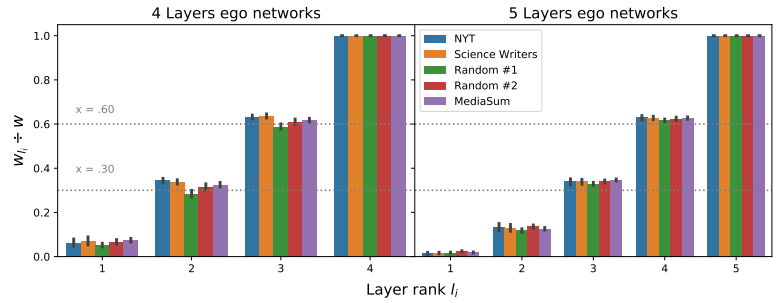
Table 10: Distribution of the optimal number of layers at each iteration of our recursive method on the Mediasum dataset. Each row contains egos with different numbers of total iterations, respectively 1, 2, and 3.

for statistical reliability. In Figure 37a, 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/X 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 6.3.5, 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 (∼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 37b, 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 $l_{\tau-2}$



(a) Plain



(b) Normalized

Figure 37: 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 11: Average ratio between a layer size w_{l_i} and the active size of the ego network w , in all datasets.

accounts for 30%:

$$\begin{cases} \frac{w_{l_{\tau-1}}}{w} \simeq 0.6 \\ \frac{w_{l_{\tau-2}}}{w} \simeq 0.3 \end{cases} \tag{20}$$

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 11). These values are very similar to those obtained in Chapter 4 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: $\frac{w_{l_i}}{w_{l_{i-1}}}$. As we can see in Figure 38 the ratio is very similar across the datasets for $i \geq 3$. The ratio tends to reach a value

slightly below two toward the outermost layers. These results are the same as those obtained in Chapter 4.

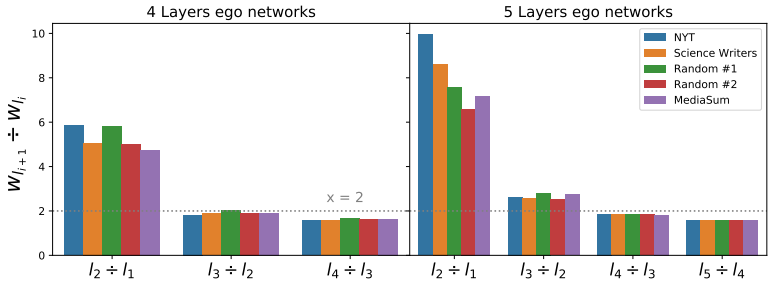


Figure 38: Scaling ratio.

When comparing the current findings with previous Chapters 4 and 5 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 surprisingly align well with the current findings, particularly concerning the number of circles (which were found to be between 5 and 7 in Chapter 4 vs 4-5 in this chapter) and the scaling ratio (approximately the same in Chapter 4). However, when examining the absolute sizes of individual layers, we notice larger sizes in this work compared to Chapter 4. To better understand this behavior, we can focus on the Twitter/X datasets, which are common to both studies (same users, shorter timelines in Chapter 4). Both the similarities and differences in the ego networks can be explained by the fact that the observed timelines in previous chapters 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 6.3.3). Note that the slightly higher number of optimal circles in previous chapters 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.

6.4.3 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 analysis 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:

$$\hat{w}^n = \begin{cases} w^n & \text{when} \quad n \in [0, n_a] \\ w^{n_a} & \text{when} \quad n \in [n_a, n_f] \end{cases} \tag{21}$$

In Figure 39 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). Figure 39 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.

6.4.4 Temporal stability of the active network size

With the methodology introduced in Section 6.3, 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

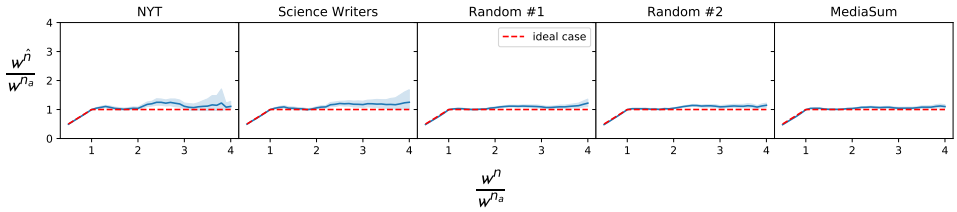


Figure 39: The stability of the algorithm is close to the ideal case.

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 40). 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 6.3.4). 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 41, 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 41 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

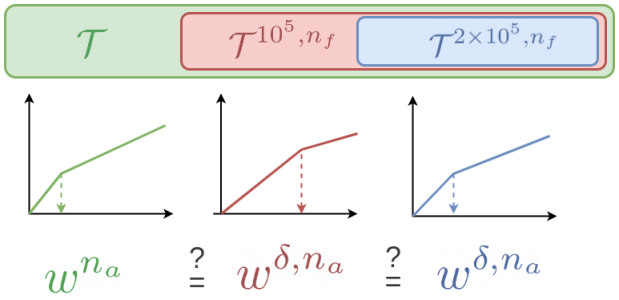


Figure 40: The diagram illustrates the temporal analysis procedure of the active size of an ego network. A temporal change corresponds here to a change in the index δ of the first word of the sequence used to build the ego network. This change leads by construction to a different saturation curve from which we will extract and study the variability of the active part size w^{δ,n_a}.

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

6.5 Conclusion

This chapter 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. It is noteworthy that these structural invariants are also present when we study oral language production, as we were able to verify with the MediaSum dataset.

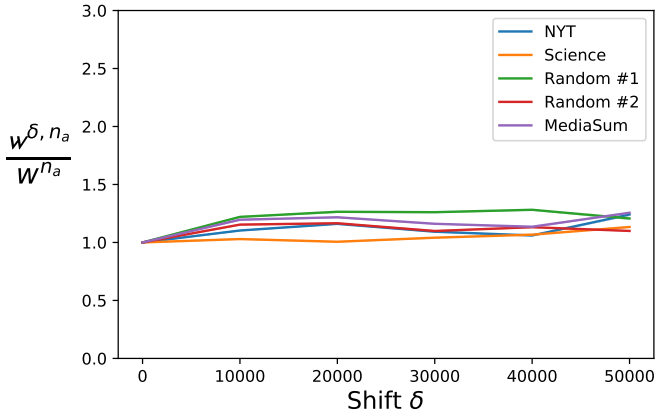


Figure 41: The shift δ of the token sequence from which the ego networks are built has almost no influence on the active size w^{δ,n_a} on. In order to average that behaviour at the dataset level, we consider the ratio $\frac{w^{\delta,n_a}}{w^{n_a}}$ where the divisor is the original active size ($\delta = 0$). 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

7 Conclusion and future work

In this thesis, we introduced the novel concept of *"ego network of words"* and offered a well-structured methodology for extracting these ego networks from textual data. This innovative approach has not only shed light on previously unexplored aspects of language but has also revealed intriguing structural and semantic invariants. These findings provide valuable insights into the shared cognitive constraints that influence the production of language.

More specifically, we investigated in Chapter 4, whether a regular structure could emerge from a frequency-based classification of the words used by a person, as a symptom of cognitive limit in the mental process. This was motivated by the fact that other mental processes are known to be driven by cognitive constraints, such as the way humans allocate cognitive capacity to social relationships. This phenomenon has been highlighted via the ego-network model, which is organized in concentric circles around the so-called "ego" and provides a "micro" perspective on the personal organization of social relationships. To conduct a similar analysis in the area of language production, we collected a diverse dataset from Twitter/X (identified as one of the major sources of informal and spontaneous language online), including tweets from regular Twitter/X users and from professional writers. After cleaning up the text by filtering out stop words and using a stemming algorithm, we applied a methodology similar to the one used to uncover social constraints: instead of grouping an ego's relationships according to their intensity, we group the words in the ego's vocabulary according to their frequency of use. These groups, whose optimal number is automatically found by the clustering algorithm, correspond, in fact, to the rings of the ego network of words, from the innermost one containing the most frequent words to the outermost one containing the least frequent words. A circle simply corresponds to the cumulative segmentation of the ego network. By studying the structure of these ego networks of words through all the users of all datasets, we uncovered regularities that constitute preliminary evidence of the aforementioned cognitive constraints. Specifically, we found that, similarly to the social case, a concentric layered structure very well captures how an individual organizes their cognitive effort in language production. Words can be grouped typically in between 5 and 7 circles, regardless of the specific class of users. We also observe a structural invariant in the size of the circles, which grow approximately 2-3 times when moving from a circle to the next one. A second structural invariant emerges for the external circles, which, regardless of the number of circles in the model, consistently account for approximately 60% and 30% of the words in the entire ego network.

Then, going beyond words as units of language, we performed in Chapter 5 a semantic analysis of the ego network of words. For this purpose, we assigned each ego network's rings a *"semantic profile"* that captures the topics associated with the words in the ring. Using a word embedding algorithm, then a clustering algorithm, we identified 100 distinct topics for each dataset. The semantic profile of a ring simply corresponds to the distribution of the words it contains in the 100 topics aforementioned. Following an in-depth analysis, we discovered that ring #1 (the innermost one which contains the most frequent words) has a special role in the model: it is semantically the most dissimilar out of all rings and also the one that generates proportionally the largest number of topics. Based on this semantic profile unique to each ring, we also separated the important (or primary) topics from the others using a simple natural breaks classification algorithm. We 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 sense, the 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 network, weaker signals show a more complex structure of influence among topics "resident" in different layers of the ego network of words.

We finally introduced 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. In practice, this means that if too many words are used to build the ego network, it ends up containing only a small number of rings (one or two), which completely undermines the relevance of the model. By studying the frequency of appearance of new words for each person, we observe that the resulting curve very often observes a two-phase behavior: first, an almost linear increase in the emergence of new words, followed by a sudden slowdown after a transition point. We deduce that the number of words included in the first phase corresponds to the number of words frequently (or actively) used by this person. We proposed a robust methodology to extract this number from the curve, which we used to define the maximum size limit for a given ego network of words. We validated this method on significantly larger datasets than in the previous two chapters: we extended the Twitter/X datasets by increasing the maximum number of tweets from 2.5K to 10K (from the same users), and we used the MediaSum dataset, which consists of

oral interviews transcripts from both CNN and NPR for which we selected the most prolific speakers. By limiting the ego network to its active part, we recovered a complex structure (between 4 and 6 circles), similar to that observed in the first two chapters, when compliance with the maximum active size was implicit because of the smaller volume of words per ego. We also proved that this maximum active size was stable over time and specific to each person, which means that it may reflect a personal cognitive limit in the use of one's vocabulary. After a further analysis, we also confirmed that the main structural properties of the active ego network of words for both the Twitter/X datasets and the MediaSum dataset are the same as those obtained in the previous two chapters. Indeed, the relative size of ego network circles are similar, and in particular the two circles before the outermost one account for consistently 30% and 60% of the words in the ego network, whatever the number of circles. The scaling ratio (when considering the circles of the ego network from the innermost one to the outermost one) is also fairly stable, since with the exception of circle 1 and 2, this ratio is around two for all datasets. This means that adding an external ring on average doubles the size of the ego network. Finally, it is noteworthy that these results were validated for both written and oral language production. This means that these two types of communicative act, each with a different physical manifestation, are subject to the same common cognitive limits, linked to the storage and use of the mental lexicon.

There are many ways to further pursue the work carried out in this thesis. The first way would be to extend the application scope of the ego network of words, to test whether other modalities of language production are subject to these cognitive constraints (and if so, to what extent). For example, one could test the model on non-English datasets, and study whether languages from the same family have a similar ego-network structure, such as the study performed in [158] with Zipf rules on various European languages. The spectrum of the communication medium can be extended by examining, for instance, discord conversations [162] or Twitch livestream chat messages [141] which correspond to new and complex patterns of conversation as a mixture of public and interpersonal messages.

A second way to extend this research work would be to use the ego network to compare the ego networks according to the authors' or interlocutors' individual properties. With the majority of the population having access to the Internet, it is becoming increasingly easy to investigate linguistic production as a function of age and socio-economic category. We could also intersect our research with that carried out on social relations and study the size of the ego network of words depending on the location of the

interlocutor in the social ego network. Interpersonal language contains social information about the level of intimacy [102] (in [129] the authors leverage levels of uncertainty and swearing to predict the familiarity between two interlocutors). In this way, an ego network of words (built for a particular conversation), could be used as a microscope to identify structural and semantic features of language in a specific context of use.

A third line of research to advance the work presented in this thesis concerns the use of the ego network of words. We have described its structural and semantic properties, but if the limits we have identified are indeed the symptoms of human cognitive constraints, do they appear in artificial language production? Due to the latest generative language models, it is now very difficult to distinguish text written by a human from text written by an AI, which opens up a whole new set of challenges in the field of bot detection [166]. Knowing that the best models produce text with statistical properties close to those produced by humans [37], our model, which is more flexible than a simple Zipf law, can potentially be used to find weak signals of suspicious patterns.

A Supporting information

A.1 Data preprocessing: filtering out inactive Twitter/X users

In order to be relevant to our work, a Twitter/X account must be an active account, which we define as an account not abandoned by its user and that tweets regularly. A Twitter/X account is considered abandoned, and we discard it, if the time since the last tweet is significantly bigger (we set this threshold at 6 months, as previously done also in [15]) than the largest period of inactivity for the account. We also consider the tweeting regularity, measured by counting the number of months where the user has been inactive. The account is tagged as sporadic, and discarded, if this number of months represents more than 50% of the observation period (defined as the time between the first tweet of a user in our dataset and the download time). We also discard accounts whose entire timeline is covered by the 3200 tweets that we are able to download, because their Twitter/X behaviour might have yet to stabilise (it is known that the tweeting activity needs a few months after an account is created to stabilise).

A.2 Ruling out soft clustering for the creation of semantic profiles

In discussed in Chapter 5, the hard clustering approach to topic extraction yields many unassigned words (Table 4). We have thus also tested soft clustering, where by each word occurrence is assigned, in any case, a probability distribution of belonging to one of the 100 topics. In Fig 42 we plot the fraction of the semantic profile covered by the top-x topics in the ring (where top-x is computed based on the semantic profile $P_r^{(e)}$). Unlike hard clustering, soft clustering gives non-zero values to the least important topics of the ring. While soft clustering allows us to include all tweets in our analysis, it has a very negative side effect. As we show in the following of the section, very generic topics become prevalent, and mask more characteristic topics that hard clustering reveals, particularly for the innermost rings. Notice that this side effect makes all rings look alike in terms of number of active topics, as we can see from the fact that all distribution curves overlap in the right-hand side plots of Fig 42.

To better investigate this aspect, we extract the important topics as described in Section 5.3.3. With two classes (important vs non-important), we obtain an average silhouette score of 0.9, confirming the good cluster

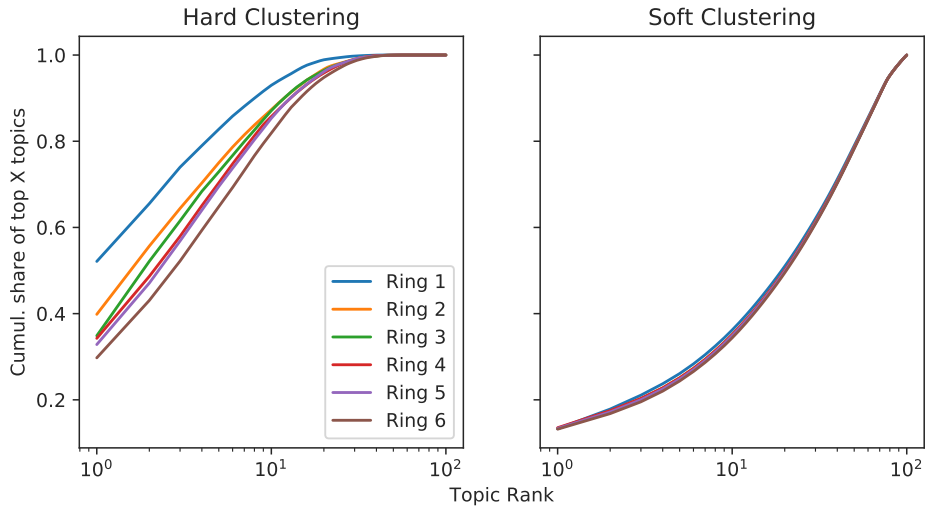


Figure 42: **Hard vs soft clustering.** Fraction of the semantic profile covered by the top-x topics in the ring, after hard (left) and soft (right) clustering.

configuration. We show these results for the Journalists dataset but similar conclusions can be drawn for the others. In Fig 43, we compare the level of importance of the 5 most dominant topics in the dataset (those who are important in the largest number of rings regardless of ego and ring rank), in the case of soft clustering and hard clustering. The figure shows that soft clustering allows some topics to dominate the whole Journalists dataset. With soft clustering, topics 93, 51, 55, 95 and 72 are important for all six rings (the ego line is filled with colored squares) of more than 50% of the ego networks. This, instead, is not the case when using hard clustering. The dominating topics in the case of soft clustering turn out being very generic ones. This is confirmed by looking at the most characteristic words in these topics in Table 14. For example topics 93 and 51, which were already among the most frequent in the hard cluster case are omnipresent in the soft cluster case, in addition to the topic 95 which is also generic but does not appear in the case of the hard cluster. We can therefore conclude that the price of a complete inclusion of tweets in our topic analysis through soft clustering only increases the noise level for all ego networks, materialized by a set of very generic topics that blur the real semantic characteristics of the rings. This is why we decided to put aside the results related to the soft clustering, in order to keep only the semantic distributions resulting from the hard

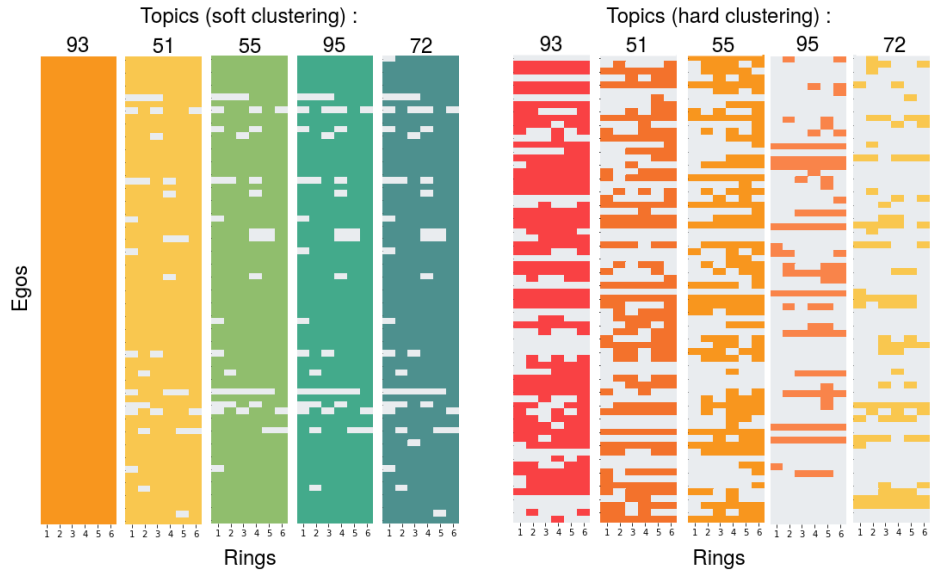


Figure 43: **Hard vs soft clustering: five most dominant topics.** The two figures show how the five most important topics in the Journalists dataset are distributed, in the case of hard clustering (on the left) and soft clustering (on the right). For each topic, a grid is drawn in which the colored square means that the corresponding topic belongs to the most important topics of ring X of the ego network Y. Those topics are important for all six rings (the line is fully colored) for respectively 49%, 28%, 19%, 21%, 9% of all the ego networks of the dataset for the hard clustered configuration (left) and 100%, 75%, 75%, 74%, 68% for the soft clustered configuration.

clustering of HDBSCAN. Note that, in light of these results, the fact that we use only a small subset of available tweets does not impact on the relevance of our analysis. What we exclude are the tweets related to "noise" topics, in the sense that they are not able to strongly characterise the Twitter/X behaviour of users, and we focus only on tweets that are strongly belonging to topics, i.e., on the semantically characteristic part of users' Twitter/X activity. A consequence of this choice is that a tweet can only be associated to one topic with hard clustering. However, if HDBSCAN assigns a tweet as an outlier, it means that it is not close enough to any topic. Otherwise, it considers the dominant topic. Finding a way to perform a finer analysis of topics without adding too much noise is an interesting direction for future work.

A.3 Additional tables

Table 12: **Hashtags, links, emojis in the datasets.** In the process of word extraction, the tweet is decomposed in tokens which are usually separated by spaces. These tokens generally corresponds to words, but they can also be links, emojis and others markers that are specific to the online language such as hashtags. The table gives the percentage of hashtags, links and emojis, which are tokens filtered out from the datasets.

	Percentage of hashtags	Percentage of links	Percentage of emojis
Journalists	1.34 %	7.27 %	0.20 %
Science writers	3.47 %	8.02 %	0.55 %
Random users #1	16.84 %	6.97 %	5.21 %
Random users #2	7.20 %	6.42 %	4.60 %

Table 13: **Example of word extraction results.**

Original tweet content	List of words after pre-processing
The @Patriots say they don't spy anymore. The @Eagles weren't taking any chances. They ran a "fake" practice before the #SuperBowl	spy, anymore, chance, run, fake, practice
#Paris attacks come 2 days before world leaders will meet in #Turkey for the G20. Will be a huge test for Turkey.	attack, come, day, world, leader, meet, huge, test, turkey
Latest garden species - the beautiful but destructive rosemary beetle, and a leafhopper (anyone know if this can be identified to species level from photo? Happy to give it a go) #30DaysWild #MyWildCity #gardening	late, garden, specie, beautiful, destructive, rosemary, beetle, leafhopper, know, identify, specie, level, photo, happy

Table 14: **Most characteristic words per topic.** They are obtained with a TF-IDF scoring.

Topic	Characteristic words (TF-IDF)
64	new obama administration tax white house comey donald president trump
24	cook lunch like dinner cheese chicken restaurant pizza food eat
93	boston read old like summer blue think google vega know
62	gop house obamacare vote repeal cut health senate republican tax
51	past february day tennis sentence week yesterday month ago year
93	boston read old like summer blue think google vega know
51	past february day tennis sentence week yesterday month ago year
55	london orleans nyc brooklyn statue monument time confederate new york
95	happy nice kind christmas great glad love thanks good thank
72	sharif judge state case pakistan gay execution supreme court arkansas

Table 15: **Topics of the NYT journalists dataset.** Most characteristic words and distribution in rings.

Topic	Characteristic words (TF-IDF)	R1	R2	R3	R4	R5	R6
0	australia australian story indigenous new	.006	.008	.009	.005	.005	.003
1	yankee baseball game pitch hit	.010	.007	.011	.011	.012	.011
2	italian soccer migrant libyan team	.009	.010	.004	.009	.004	.006
3	alabama governor senate robert moore	.000	.001	.001	.003	.004	.005
4	horse derby kentucky win race	.000	.002	.002	.003	.001	.004
5	apple mac use new silver	.001	.003	.009	.005	.005	.006
6	midwest south city today times	.013	.010	.010	.012	.014	.010
7	fox news pope fake vatican	.015	.009	.010	.006	.006	.008
8	french election macron pen paris	.000	.000	.008	.004	.002	.001
9	white shark nationalist president harvard	.002	.003	.003	.005	.003	.003
10	black slave african american asian	.014	.024	.021	.021	.022	.018
11	turkey turkish referendum protester president	.001	.003	.004	.001	.002	.001
12	cat mouse kitten game bureau	.001	.004	.006	.002	.004	.005
13	birthday happy halloween spring valentine	.002	.001	.001	.003	.000	.001
14	sleep bed nap asleep bedtime	.003	.004	.006	.007	.012	.007
15	phone sorry storm stuck quick	.006	.014	.006	.007	.010	.009
16	german right english angela fluent	.000	.001	.001	.000	.001	.001
17	football bowl super player anthem	.004	.007	.008	.005	.006	.007
18	brazil president brazilian scandal rio	.000	.003	.002	.003	.002	.004
19	flight plane fly helicopter passenger	.002	.002	.002	.001	.002	.002
20	beer vest clock declare power	.005	.006	.006	.009	.008	.005
21	dog pet puppy love good	.004	.005	.006	.006	.005	.005
22	wine red carpet school good	.007	.008	.008	.007	.006	.004
23	fish boat surf fishing sea	.000	.000	.002	.001	.005	.001
24	eat food pizza restaurant chicken	.001	.003	.004	.005	.008	.006
25	train subway station new delay	.000	.003	.001	.003	.002	.004
26	canada canadian refugee indigenous new	.001	.001	.002	.001	.001	.002
27	year minute yahoo day hour	.014	.013	.011	.009	.006	.005
28	bear montana wolf colorado wood	.003	.006	.003	.004	.004	.004
29	hockey game team stanley cup	.000	.002	.001	.000	.001	.000
30	snow ice winter cold arctic	.005	.005	.005	.003	.006	.008
31	texas special state education cap	.008	.004	.005	.007	.006	.006
32	sunday saturday night morning monday	.002	.002	.003	.003	.003	.004
33	friday thursday tuesday monday wednesday	.002	.003	.016	.006	.006	.007
34	moon space alien planet earth	.006	.004	.004	.006	.004	.005
35	japan abe japanese reactor scandal	.003	.005	.013	.008	.010	.011
36	china chinese hong new robot	.016	.007	.006	.006	.003	.007
37	north missile korean nuclear south	.000	.000	.002	.001	.001	.001
38	basketball league source trade season	.004	.006	.004	.002	.004	.003
39	sigh mike right wow know	.005	.009	.005	.004	.003	.003
40	twitter social medium like live	.003	.002	.003	.002	.002	.003
41	miss destroyer sailor collision ship	.041	.048	.041	.039	.055	.046
42	day july today year hour	.006	.006	.006	.007	.003	.003
43	movie watch film play episode	.001	.001	.001	.001	.001	.001
44	lobbyist intend dislike implication apology	.006	.016	.002	.002	.005	.002
45	california earthquake san francisco quake	.009	.003	.002	.001	.000	.000
46	hurricane florida irma storm harvey	.030	.046	.054	.069	.061	.066
47	prince woman crown ebony ballroom	.015	.007	.010	.001	.006	.004
48	iran iranian deal nuke president	.008	.004	.006	.005	.005	.005
49	syrian attack chemical strike weapon	.001	.003	.013	.008	.005	.010
50	russian russia trump investigation election	.001	.000	.002	.000	.002	.001

51	year ago month yesterday week	.009	.016	.013	.017	.010	.015
52	climate trump change paris cut	.001	.002	.001	.002	.008	.003
53	climate change oil paris carbon	.004	.002	.002	.001	.002	.002
54	tweet chronological good evergreen great	.038	.018	.037	.039	.030	.024
55	york new confederate time monument	.003	.002	.003	.003	.004	.008
56	tax estate cash bank fund	.001	.004	.001	.003	.003	.002
57	famine south yemen cholera venezuelan	.197	.147	.136	.121	.113	.127
58	listen book talk daily new	.022	.009	.003	.014	.013	.014
59	morning tomorrow good trial page	.007	.008	.006	.009	.011	.008
60	week hour month year marathon	.010	.008	.010	.007	.008	.010
61	million year billion spend marijuana	.004	.007	.006	.007	.007	.005
62	tax republican senate health cut	.023	.017	.017	.025	.031	.021
63	school high homework student college	.007	.008	.006	.006	.010	.008
64	trump president donald comey house	.003	.005	.003	.002	.002	.003
65	big palestinian time read story	.005	.002	.003	.006	.002	.005
66	fashion week mother wear model	.006	.004	.008	.005	.005	.006
67	dress leather pink skirt gown	.022	.022	.025	.026	.023	.029
68	tonight weekend atlanta bachelor georgia	.000	.001	.001	.001	.001	.001
69	song hip rap rock hop	.005	.001	.001	.000	.000	.001
70	broadway opera theater classical music	.006	.012	.002	.002	.002	.003
71	dot reporter peer time	.026	.017	.011	.021	.016	.017
72	arkansas court supreme execution gay	.006	.012	.010	.012	.009	.010
73	sexual harassment woman accuse allegation	.024	.034	.039	.024	.016	.025
74	wait bus happen mean depend	.038	.010	.007	.003	.003	.002
75	send address question shoot reach	.000	.000	.000	.001	.001	.001
76	car driver drive driving self	.001	.003	.001	.001	.002	.001
77	eclipse solar total delete totality	.029	.038	.040	.036	.037	.037
78	story news journalist accuse public	.070	.071	.084	.085	.115	.104
79	suicide trial roy conrad carter	.001	.003	.005	.005	.005	.006
80	die dead york robert roger	.001	.001	.005	.008	.004	.003
81	roe squeamish lisa susan collins	.002	.023	.005	.004	.005	.004
82	dislike unintended implication apology culture	.007	.012	.010	.011	.009	.008
83	lady girl yes elizabeth finale	.003	.004	.010	.008	.007	.007
84	book soon read write editor	.003	.000	.000	.000	.000	.002
85	best great video love game	.006	.012	.008	.010	.010	.012
86	bad terrible hate sorry awful	.002	.005	.008	.005	.005	.006
87	drug police arrest jail gang	.020	.016	.016	.014	.020	.020
88	kill militant police army congo	.000	.003	.007	.005	.003	.005
89	agree tweet important fascinate interesting	.001	.002	.002	.003	.002	.003
90	wrong bad argue moly mean	.003	.005	.011	.011	.007	.010
91	love woman genius happy sandra	.004	.005	.005	.005	.002	.004
92	yes true right joke correct	.002	.001	.002	.004	.004	.003
93	know vega google think blue	.011	.012	.011	.011	.015	.016
94	beautiful great cool gorgeous fun	.010	.012	.005	.004	.008	.006
95	good love glad great christmas	.024	.030	.015	.023	.019	.019
96	god know exactly gold yes	.016	.013	.012	.009	.010	.008
97	tho alex come like pat	.000	.002	.002	.003	.003	.002
98	kate congratulation diane karen welcome	.018	.012	.014	.028	.016	.017
99	read share contact matt paul	.003	.003	.003	.006	.003	.002

Table 16: **Topics of the science writers dataset.** Most characteristic words and distribution in rings.

Topic	Characteristic words (TF-IDF)	R1	R2	R3	R4	R5	R6
0	daily late luck today	.002	.002	.002	.003	.000	.003
1	baseball game lacrosse football player	.008	.008	.006	.007	.009	.011
2	follower week new canada right	.014	.013	.011	.017	.013	.016
3	video subtitle individual anonymous credit	.009	.005	.006	.002	.007	.005
4	aku morning good river countryside	.000	.000	.000	.000	.000	.000
5	aku morning good lake photo	.010	.004	.002	.004	.005	.002
6	badge earn level middle road	.009	.022	.017	.009	.006	.005
7	web nature post life plastic	.000	.000	.000	.000	.000	.000
8	essay environmental educator nature conservation	.009	.008	.004	.006	.007	.006
9	daily late clow soon hourly	.003	.003	.002	.005	.003	.003
10	submission album cheer shoot hello	.018	.008	.014	.012	.019	.023
11	submission album cheer shoot hello	.003	.008	.011	.007	.008	.008
12	poker play chess player best	.011	.006	.006	.006	.005	.006
13	robot human killer new job	.010	.008	.008	.005	.006	.007
14	year gorilla monkey story ape	.003	.008	.008	.007	.006	.007
15	white male quote diversity cause	.004	.010	.010	.016	.015	.016
16	christmas holiday year tree festive	.034	.049	.033	.045	.044	.056
17	plane flight fly spy airplane	.001	.005	.003	.005	.003	.004
18	eclipse space moon earth solar	.004	.002	.003	.002	.003	.003
19	african ancient beard genome revisit	.002	.003	.004	.008	.008	.005
20	air asthma pollution risk city	.003	.002	.002	.002	.002	.002
21	coffee shop drink caffeine cup	.024	.029	.022	.047	.029	.031
22	drink beer brewery beach ale	.001	.000	.000	.000	.000	.000
23	china chinese european scientific british	.003	.006	.003	.003	.002	.002
24	morning good perambulation wake bob	.003	.007	.009	.006	.011	.007
25	week virology new wildlife picture	.000	.000	.001	.000	.000	.001
26	negotiation britain tax british european	.013	.017	.007	.009	.005	.005
27	car driving self auto test	.046	.038	.038	.025	.027	.026
28	happy birthday year mother wedding	.004	.004	.008	.005	.004	.008
29	twitter mention reach social medium	.003	.004	.003	.003	.004	.003
30	apple mobile search phone new	.008	.005	.006	.007	.005	.009
31	weekly microbiology science episode new	.000	.001	.000	.001	.000	.000
32	social medium fake news combat	.011	.012	.011	.008	.009	.007
33	record hot year high warm	.039	.029	.043	.056	.050	.056
34	journalist join hear sally tonight	.007	.010	.007	.012	.009	.011
35	prize chemistry win medicine physiology	.006	.005	.008	.005	.005	.006
36	chicken meat eat animal barn	.000	.004	.003	.007	.004	.004
37	sleep bed night nap dream	.034	.036	.013	.015	.015	.018
38	earthquake quake tsunami seismic big	.003	.006	.009	.005	.005	.005
39	ice arctic winter snow antarctica	.005	.011	.009	.007	.007	.008
40	canada canadian maple citizenship government	.005	.005	.012	.011	.006	.009
41	california wildfire northern burn flee	.004	.013	.006	.005	.009	.008
42	hurricane storm flood rain irma	.006	.008	.012	.009	.009	.010
43	old fossil human year ancient	.001	.004	.005	.004	.007	.005
44	frog otter snake amphibian rid	.069	.067	.071	.067	.076	.075
45	pterosaur skull crest cornified animal	.005	.002	.003	.004	.003	.001
46	bird spider bat flower moth	.004	.009	.014	.011	.012	.015
47	dinosaur fossil bird mammal discover	.119	.107	.146	.137	.148	.142
48	shark whale sea fish ocean	.002	.002	.005	.004	.004	.006
49	bear wolf polar kill rhino	.006	.007	.008	.008	.010	.011
50	dog puppy good breed love	.002	.002	.003	.002	.001	.001

51	chocolate eat pizza pie cheese	.002	.003	.003	.002	.002	.003
52	food delicious fortune restaurant love	.009	.002	.003	.003	.002	.002
53	cat dog kitten like think	.037	.039	.032	.032	.035	.035
54	rule tobacco regulatory million health	.000	.000	.000	.000	.000	.000
55	year time hour paper china	.008	.007	.007	.007	.004	.002
56	woman award stem girl winner	.001	.002	.003	.003	.002	.004
57	editor story wired write business	.041	.033	.043	.044	.038	.041
58	car bicycle bike crash driving	.001	.002	.002	.001	.004	.002
59	solar power wind energy electricity	.000	.000	.000	.000	.000	.000
60	american america black prescription slavery	.001	.001	.005	.003	.004	.006
61	die child woman bad parent	.020	.011	.008	.009	.013	.011
62	health medical care patient doctor	.009	.005	.014	.007	.007	.010
63	photo pic sharpen color apply	.004	.005	.007	.007	.007	.008
64	cancer new cell mouse disease	.004	.002	.001	.003	.008	.005
65	chromosome human horse embryo gene	.003	.005	.006	.004	.003	.003
66	republican senate house senator white	.017	.008	.008	.009	.007	.006
67	year day week halloween time	.002	.000	.000	.002	.001	.000
68	trump administration president climate donald	.002	.005	.002	.004	.003	.003
69	nuclear north weapon war iran	.003	.006	.001	.003	.003	.002
70	coal oil climate fuel kentucky	.004	.004	.012	.005	.012	.006
71	climate change carbon scientist report	.007	.003	.003	.003	.004	.005
72	defense arrive plant episode week	.001	.008	.009	.005	.007	.006
73	kill police murder arrest officer	.010	.009	.011	.012	.009	.007
74	documentary film watch new series	.013	.014	.003	.005	.002	.001
75	year end hour ago chronicle	.000	.000	.000	.000	.000	.000
76	send address dot touch chat	.004	.002	.001	.003	.002	.004
77	science donation match great recur	.009	.003	.004	.005	.006	.008
78	like good way think know	.001	.004	.002	.003	.003	.001
79	boston stereo arena queen wed	.021	.016	.008	.010	.009	.008
80	great year sing night happy	.000	.001	.007	.003	.006	.001
81	night stream miss catch tonight	.005	.004	.006	.005	.005	.004
82	week month year new tomorrow	.001	.005	.005	.004	.004	.003
83	science student school week scientist	.006	.020	.011	.013	.019	.010
84	sunday saturday night come need	.053	.041	.044	.032	.037	.029
85	thursday friday join wednesday tuesday	.001	.001	.002	.002	.002	.002
86	science sexual harassment obituary journalism	.011	.011	.005	.008	.007	.006
87	community follow rank step work	.010	.007	.012	.010	.009	.012
88	free article site tweet want	.011	.009	.008	.006	.009	.006
89	book read weekend science journal	.000	.000	.000	.000	.000	.000
90	mean think worry thing point	.000	.004	.001	.002	.004	.004
91	know right sure check want	.012	.020	.008	.010	.006	.009
92	bad people medium crazy like	.003	.001	.001	.001	.001	.001
93	sorry bad terrible sad weird	.038	.023	.023	.031	.021	.020
94	god nope test idea know	.015	.014	.011	.009	.012	.009
95	yes agree wow mean whoa	.005	.007	.008	.012	.009	.011
96	glad kind great love enjoy	.000	.004	.003	.006	.001	.001
97	good awesome love cool nice	.002	.001	.003	.002	.003	.002
98	fan week big congratulation mull	.000	.000	.000	.000	.000	.000
99	bless andy congratulation paul mate	.003	.003	.004	.003	.003	.002

Table 17: **Topics of the random users #1 dataset.** Most characteristic words and distribution in rings.

Topic	Characteristic words (TF-IDF)	R1	R2	R3	R4	R5	R6
0	twitter mention reach week like	.003	.002	.002	.002	.003	.002
1	automatically unfollowed check follow people	.005	.004	.008	.004	.005	.004
2	natural naturally soon tune launch	.004	.005	.005	.004	.005	.005
3	post photo atlantic raw valley	.011	.012	.012	.015	.017	.017
4	week fan big boy great	.001	.001	.001	.001	.001	.001
5	replacement screen ram core battery	.005	.004	.005	.005	.005	.006
6	practice spanish read news post	.007	.007	.007	.007	.008	.008
7	bristol story chronicle daily include	.001	.001	.001	.001	.001	.001
8	cannabis marijuana medical weed industry	.002	.002	.002	.003	.004	.003
9	australia visa immigration australian apply	.006	.002	.002	.001	.001	.001
10	alert trance dance hit triple	.007	.007	.007	.005	.005	.005
11	hire job post pro apply	.000	.002	.001	.002	.004	.004
12	music available game prophesy gospel	.000	.002	.002	.002	.002	.002
13	happy peep good thanksgiving holiday	.022	.025	.029	.028	.026	.025
14	canada immigration apply express entry	.011	.013	.006	.004	.005	.006
15	visit information weekly clue chat	.001	.003	.001	.001	.001	.001
16	track rock follower today outlaw	.001	.002	.002	.003	.004	.004
17	late daily innovative horse source	.005	.004	.004	.005	.004	.004
18	moon space mar astronaut mission	.009	.004	.004	.002	.002	.001
19	road gold world win champ	.021	.017	.013	.012	.008	.009
20	link subscribe click channel registration	.003	.005	.005	.005	.004	.005
21	red blue sugar mug titan	.003	.003	.004	.004	.005	.004
22	catholic priest pope church prayer	.000	.000	.001	.001	.001	.001
23	trading risky suitable net close	.003	.002	.003	.003	.003	.005
24	life sunday breath coach insurance	.004	.007	.005	.005	.005	.004
25	god lord jesus christ unto	.005	.003	.003	.003	.002	.002
26	associate page log principal excerpt	.001	.002	.003	.002	.002	.003
27	amazon offer bank discount author	.029	.020	.017	.012	.014	.011
28	phone car tune today mobile	.006	.003	.007	.008	.009	.011
29	car hire plate vat drive	.006	.008	.011	.007	.008	.009
30	christmas merry gift festive day	.001	.003	.003	.003	.005	.005
31	friday weekend happy day halloween	.001	.001	.001	.001	.002	.002
32	tea beer drink come brewery	.001	.001	.001	.001	.002	.001
33	yoga teacher japanese meditation training	.006	.006	.007	.006	.007	.006
34	life weight lose people think	.017	.016	.016	.018	.018	.019
35	black white american fear legging	.008	.010	.009	.008	.009	.008
36	today evangelist shower angela help	.009	.009	.012	.013	.013	.014
37	thing dream life right time	.020	.028	.026	.027	.029	.029
38	monday week morning happy good	.004	.006	.009	.007	.007	.007
39	coffee cup morning good day	.009	.017	.009	.007	.005	.005
40	password best wednesday frustration day	.004	.003	.004	.004	.004	.003
41	dog pet puppy love dane	.006	.003	.004	.004	.005	.004
42	cat kitten home lover happy	.009	.008	.011	.013	.013	.014
43	apply badge level earn job	.019	.006	.006	.004	.003	.003
44	tuesday today day good life	.004	.006	.005	.006	.005	.005
45	food breakfast eat recipe chris	.013	.017	.015	.019	.022	.021
46	cake chocolate cream ice birthday	.006	.006	.009	.006	.006	.008
47	look nice delicious yummy forward	.006	.005	.008	.008	.009	.011
48	flight dana fly update gate	.009	.008	.009	.010	.008	.010
49	chicken curry lunch green menu	.005	.005	.003	.004	.003	.004
50	follow hey kindly smile fib	.010	.009	.011	.013	.011	.010

51	bedroom home house pool village	.031	.032	.036	.035	.034	.034
52	shop fashion dress wedding buy	.008	.008	.006	.005	.007	.005
53	cricket win match wicket cup	.005	.004	.004	.005	.004	.004
54	win rocket game final score	.001	.001	.001	.001	.001	.001
55	basketball football team game soccer	.055	.051	.056	.055	.053	.057
56	beautiful cute hope bird look	.004	.001	.001	.001	.001	.001
57	sorry inconvenience contact hear team	.006	.006	.008	.008	.006	.007
58	sleep bed night wake nap	.005	.006	.006	.006	.007	.005
59	winter snow cold ski rain	.002	.001	.002	.003	.002	.002
60	tonight winner night win ticket	.013	.011	.008	.009	.006	.006
61	connect let follow group family	.008	.009	.009	.010	.011	.011
62	social medium hilarious engagement marketing	.010	.004	.005	.003	.002	.002
63	live music official video bad	.022	.024	.025	.027	.029	.025
64	dance befit class studio join	.001	.001	.001	.002	.002	.002
65	video learn color alphabet child	.006	.006	.006	.006	.005	.006
66	content write writer currently start	.011	.010	.009	.009	.009	.007
67	climate east change late south	.057	.038	.036	.033	.033	.035
68	stay park hostel hotel board	.015	.022	.019	.024	.020	.019
69	oil climate join fossil fuel	.089	.103	.104	.106	.098	.093
70	birthday happy wish bless year	.009	.006	.009	.006	.007	.006
71	morning good golf bless day	.008	.009	.008	.008	.010	.008
72	address send hello look certainly	.001	.001	.001	.001	.001	.002
73	staff dudley health nurse mental	.008	.010	.011	.011	.010	.011
74	vulnerable rat outstanding agency child	.002	.002	.004	.005	.007	.006
75	help miss autism interested locate	.004	.006	.003	.003	.003	.003
76	tutor tip directory literacy foot	.003	.001	.001	.002	.001	.001
77	cancer patient therapy cell treatment	.006	.003	.000	.000	.000	.000
78	west movie blast film watch	.023	.022	.026	.026	.026	.024
79	million year store billion investment	.001	.002	.001	.001	.001	.001
80	new salary happy year profile	.002	.001	.002	.003	.002	.002
81	appreciate share shout homeless tweet	.003	.002	.002	.002	.002	.002
82	school exam dismissal free generate	.016	.021	.019	.019	.018	.017
83	mother brother son queen love	.011	.012	.013	.015	.016	.017
84	book savvy silly society update	.004	.004	.004	.004	.005	.005
85	woman ass sexy sensual sophisticated	.015	.018	.018	.018	.014	.015
86	day verse valentine great grateful	.002	.002	.003	.003	.001	.002
87	cloud marketing digital network robot	.106	.105	.104	.107	.111	.111
88	career business support information employer	.001	.003	.002	.002	.001	.002
89	year wait month code week	.006	.009	.007	.005	.005	.004
90	welcome sacrifice salute nancy champagne	.009	.010	.014	.015	.015	.016
91	love congratulation feedback great hug	.003	.005	.005	.004	.004	.004
92	creation awesome create think look	.018	.019	.020	.021	.024	.025
93	india anniversary indian birth kashmiri	.017	.013	.009	.008	.007	.008
94	amen preach word naa ouch	.004	.004	.005	.005	.006	.006
95	dream true agree old believe	.003	.003	.003	.005	.005	.004
96	vote know people hold yes	.006	.009	.008	.008	.011	.012
97	trump president russia hillary lawyer	.007	.004	.007	.006	.006	.005
98	arrest police man kill murder	.000	.001	.001	.002	.001	.002
99	bad sad disrespectful awful disgust	.000	.000	.000	.000	.000	.000

Table 18: **Topics of the random users #2 dataset.** Most characteristic words and distribution in rings.

Topic	Characteristic words (TF-IDF)	R1	R2	R3	R4	R5	R6
0	temp sea pressure rain weather	.008	.009	.012	.012	.011	.011
1	job check nurse advisor ref	.002	.004	.006	.003	.006	.006
2	data storage file holiday song	.004	.007	.011	.007	.006	.007
3	morning good kevin steve vacancy	.019	.013	.015	.018	.014	.014
4	live saturday stream masquerade laugh	.011	.010	.009	.008	.007	.004
5	jump long pit radio runway	.002	.005	.004	.005	.002	.002
6	pitch synthetic turf artificial sport	.006	.008	.006	.008	.009	.007
7	market sign september risk easy	.000	.000	.000	.000	.000	.000
8	consultant resin sport pitch flooring	.009	.012	.013	.008	.007	.008
9	aquarius sensational today seventy happen	.017	.021	.019	.016	.019	.022
10	playground marking key stage game	.012	.013	.009	.017	.011	.013
11	gallery collection art contemporary home	.006	.002	.002	.003	.002	.002
12	cancer mouth breast research today	.005	.008	.010	.007	.008	.008
13	safety train air cylinder pneumatic	.019	.021	.019	.020	.020	.017
14	trade wale choose big car	.006	.005	.004	.005	.004	.006
15	beautiful cute look amaze adorable	.066	.057	.063	.077	.072	.075
16	manager director yoga technical executive	.008	.011	.007	.006	.008	.011
17	course training certificate lunch click	.005	.007	.004	.004	.004	.005
18	news north northern west warrior	.040	.042	.045	.045	.050	.043
19	mobility product salary showroom look	.002	.002	.002	.002	.003	.003
20	china chinese outbreak congo measles	.022	.021	.022	.023	.023	.021
21	ref level surfacing sale representative	.009	.010	.013	.009	.009	.007
22	interview job excellent benefit tip	.001	.002	.003	.003	.002	.002
23	truck law year new minute	.015	.017	.013	.011	.008	.008
24	privacy place data speaker security	.011	.009	.010	.011	.012	.011
25	rule entry voucher submit year	.007	.008	.007	.005	.007	.006
26	late daily predator bullet pip	.001	.002	.002	.003	.003	.003
27	twitter mention reach week like	.019	.018	.020	.021	.022	.018
28	birthday happy hope soon wish	.003	.003	.004	.005	.005	.006
29	cheer agree true baby mate	.004	.005	.004	.004	.004	.004
30	hockey court tennis final surface	.003	.002	.002	.002	.004	.002
31	free instant horse audit tip	.001	.000	.000	.000	.001	.000
32	branch rate store available buy	.000	.000	.000	.000	.000	.000
33	support help child people sport	.011	.017	.013	.010	.014	.014
34	tropical storm thunderstorm weather rain	.010	.009	.009	.010	.012	.009
35	sleep bed night nap asleep	.003	.002	.002	.004	.002	.002
36	property bedroom tax station family	.018	.011	.005	.004	.004	.005
37	ship cruise new marine boat	.019	.028	.029	.033	.037	.033
38	new star review unit charge	.003	.004	.003	.003	.003	.004
39	wine competition enter medal sommelier	.004	.004	.003	.006	.004	.005
40	cost value low decision help	.007	.007	.008	.010	.009	.007
41	coffee tea cup lunch grandma	.003	.004	.003	.004	.004	.005
42	christmas merry gift year festive	.047	.041	.052	.054	.042	.051
43	click link workshop business poetry	.040	.040	.043	.040	.046	.042
44	night tonight bar drink beer	.006	.005	.006	.006	.007	.006
45	number guide model mary information	.033	.022	.015	.016	.010	.012
46	cat kitten bruce love like	.049	.037	.053	.052	.054	.057
47	food farm production course eat	.002	.005	.001	.001	.002	.001
48	attack data breach security user	.024	.039	.043	.041	.038	.035
49	garden summer plant grow flower	.005	.004	.005	.007	.008	.008

50	dog puppy pet guide animal	.007	.006	.003	.004	.003	.005
51	pizza chicken cheese meat sausage	.004	.005	.005	.004	.005	.004
52	miss today gemini watch courtesy	.007	.004	.007	.008	.006	.007
53	health mental cigarette tobacco cricket	.001	.001	.002	.002	.001	.001
54	brain injury scientist researcher science	.000	.002	.001	.000	.001	.001
55	climate green change carbon environmental	.022	.009	.009	.005	.007	.006
56	fisherman fish beanie fishery marine	.005	.001	.002	.001	.002	.001
57	photo learn range support publish	.015	.018	.018	.016	.019	.017
58	follow automatically unfollowed check person	.001	.003	.002	.002	.002	.002
59	movie game best funny hot	.001	.000	.001	.001	.001	.000
60	happy car customer new trade	.004	.004	.004	.006	.005	.005
61	shower black halloween dance look	.010	.011	.010	.010	.012	.013
62	ray order edition release win	.026	.029	.028	.030	.030	.035
63	tomorrow evening close support message	.006	.005	.005	.007	.004	.005
64	ticket tour sale wait announce	.002	.006	.008	.007	.006	.006
65	music album new single pic	.011	.011	.010	.011	.012	.013
66	monday wednesday tuesday flight fly	.021	.018	.019	.022	.019	.023
67	friday library thursday fact hub	.005	.008	.009	.007	.008	.007
68	pisces scorpio virgo aries stop	.004	.005	.004	.004	.005	.006
69	rugby story world news cup	.012	.008	.007	.006	.007	.007
70	red player play win game	.003	.003	.004	.002	.002	.002
71	football league weekend win round	.010	.011	.012	.012	.013	.014
72	model age commercial shoot female	.014	.013	.008	.005	.004	.004
73	trump donald president like america	.002	.005	.006	.006	.008	.008
74	school bullying start change cover	.005	.010	.009	.007	.010	.009
75	student need require math support	.002	.003	.002	.002	.002	.003
76	congratulation award woman queen category	.006	.007	.007	.008	.007	.006
77	sorry order address number hear	.005	.002	.002	.002	.001	.001
78	update android store form creator	.005	.005	.005	.005	.008	.006
79	day valentine today good happy	.019	.014	.013	.012	.013	.010
80	cement airport retail duty travel	.002	.002	.002	.002	.003	.002
81	month year week contract sunday	.006	.012	.009	.008	.007	.005
82	business parent feature social medium	.003	.004	.003	.002	.002	.002
83	bank payment launch platform banking	.009	.008	.006	.006	.007	.006
84	week hour month image shot	.017	.017	.015	.020	.016	.018
85	leadership development network skill leader	.003	.005	.005	.005	.005	.003
86	hug paw love send david	.003	.005	.003	.003	.004	.003
87	police man old jail arrest	.012	.012	.010	.010	.010	.010
88	road car driver vehicle cyclist	.012	.010	.008	.008	.008	.008
89	car race drive raceway driver	.017	.014	.013	.009	.011	.006
90	vote boris deal labour party	.001	.002	.002	.002	.003	.004
91	tip time management try start	.006	.010	.011	.007	.009	.016
92	address send password congratulation number	.007	.010	.009	.007	.007	.008
93	bad hate sad sorry wrong	.004	.005	.005	.005	.006	.006
94	subscription address sorry hear look	.001	.001	.002	.001	.001	.002
95	echo team sorry touch order	.003	.002	.002	.002	.004	.004
96	dont think know game like	.011	.011	.005	.006	.004	.005
97	love sun island change book	.002	.003	.003	.004	.003	.005
98	big follower fan week share	.010	.005	.006	.006	.007	.007
99	yes yeah xmas amen everyday	.007	.005	.009	.007	.006	.004

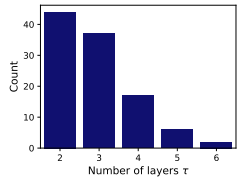
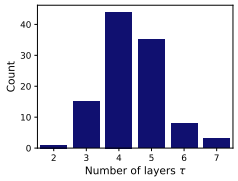
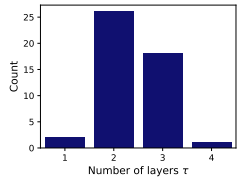
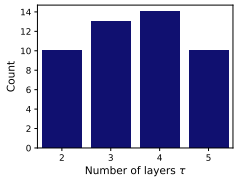
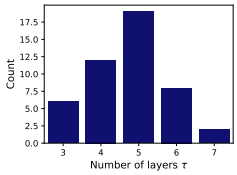
No cut	1st iteration	2nd iteration
		
		

Table 19: 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.

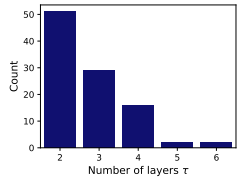
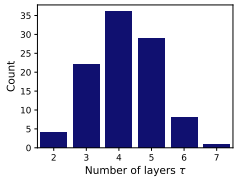
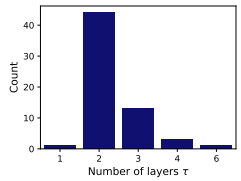
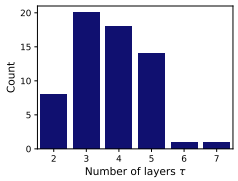
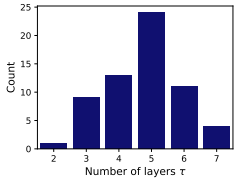
No cut	1st iteration	2nd iteration
		
		

Table 20: 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.

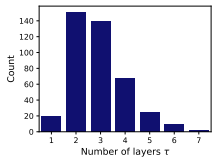
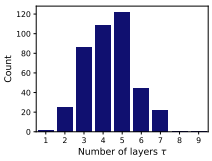
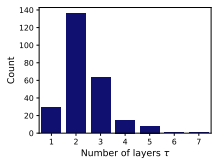
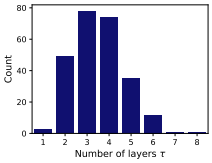
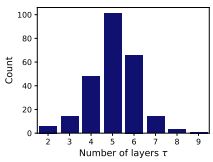
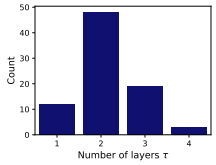
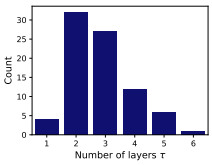
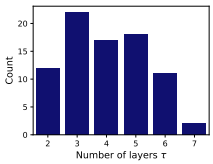
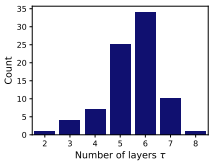
No cut	1st iteration	2nd iteration	3rd iteration
			
			
			

Table 21: 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.

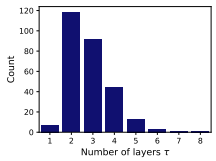
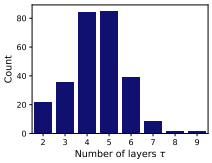
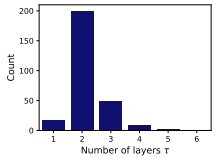
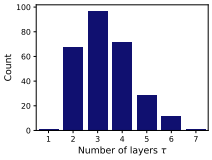
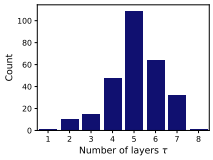
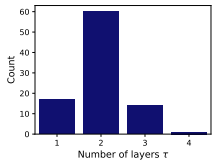
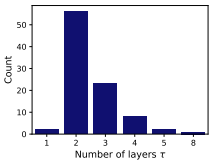
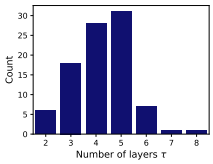
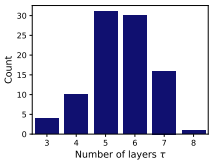
No cut	1st iteration	2nd iteration	3rd iteration
			
			
			

Table 22: 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.

B Bibliography

References

[1] Bilal Abu-Salih, Pornpit Wongthongtham, and Kit Yan Chan. Twitter mining for ontology-based domain discovery incorporating machine learning. *Journal of Knowledge Management*, 2018.

[2] Ralph Adolphs. Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3):165–178, 2003.

[3] Ralph Adolphs. The biology of fear. *Current biology*, 23(2):R79–R93, 2013.

[4] Sinan Aral and Marshall Van Alstyne. The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1):90–171, 2011.

[5] Valerio Arnaboldi, Marco Conti, Massimiliano La Gala, Andrea Passarella, and Fabio Pezzoni. Information diffusion in OSNs: the impact of nodes' sociality. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 616–621. ACM, 2014.

[6] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin Dunbar. Dynamics of personal social relationships in online social networks: a study on twitter. In *Proceedings of the first ACM conference on Online social networks*, pages 15–26. ACM, 2013.

[7] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin IM Dunbar. Online social networks and information diffusion: The role of ego networks. *Online Social Networks and Media*, 1:44–55, 2017.

[8] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. 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*, pages 31–40. IEEE, 2012.

[9] Valerio Arnaboldi, Robin IM Dunbar, Andrea Passarella, and Marco Conti. Analysis of co-authorship ego networks. In *International Conference and School on Network Science*, pages 82–96. Springer, Springer, Cham, 2016.

[10] Valerio Arnaboldi, Andrea Passarella, Marco Conti, and Robin Dunbar. Structure of ego-alter relationships of politicians in twitter. *Journal of Computer-Mediated Communication*, 22(5):231–247, sep 2017.

[11] Michael S Ayers and Lynne M Reder. A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review*, 5(1):1–21, 1998.

[12] Vijay Balasubramanian. Brain power. *Proceedings of the National Academy of Sciences*, 118(32):e2107022118, 2021.

[13] Buster Benson. Cognitive bias cheat sheet. *Better Humans*, 2016.

[14] Chris Bentz and Ramon Ferrer Cancho. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, pages 1–4. University of Tübingen, 2016.

[15] Chiara Boldrini, Mustafa Toprak, Marco Conti, and Andrea Passarella. Twitter and the press: an ego-centred analysis. In *Companion Proceedings of the The Web Conference'18*, pages 1471–1478, 2018.

[16] Donald E Broadbent. Word-frequency effect and response bias. *Psychological review*, 74(1):1, 1967.

[17] Marc Brysbaert, Paweł Mandera, and Emmanuel Keuleers. The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27(1):45–50, 2018.

[18] Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 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*, 7(JUL):1116, jul 2016.

[19] Turhan Canli, Heidi Sivers, Susan L Whitfield, Ian H Gotlib, and John DE Gabrieli. Amygdala response to happy faces as a function of extraversion. *Science*, 296(5576):2191–2191, 2002.

[20] Alfonso Caramazza. How many levels of processing are there in lexical access? *Cognitive neuropsychology*, 14(1):177–208, 1997.

[21] Peter Carruthers. The cognitive functions of language. *Behavioral and brain sciences*, 25(6):657–674, 2002.

[22] Logan Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Z Dara. Strolovitch (2017),"intertemporal differences among mturk worker demographics,". Technical report, Working Paper, University of Michigan.

[23] Valerie M Chase, Ralph Hertwig, and Gerd Gigerenzer. Visions of rationality. *Trends in cognitive sciences*, 2(6):206–214, 1998.

[24] Michelene TH Chi. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, 1(1):73–105, 2009.

[25] Noam Chomsky. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15, 1980.

[26] Noam Chomsky. Three factors in language design. *Linguistic inquiry*, 36(1):1–22, 2005.

[27] Morten H Christiansen and Nick Chater. Language as shaped by the brain. *Behavioral and brain sciences*, 31(5):489–509, 2008.

[28] Benjamin A Clegg, Gregory J DiGirolamo, and Steven W Keele. Sequence learning. *Trends in cognitive sciences*, 2(8):275–281, 1998.

[29] Guillem Collell and Jordi Fauquet. Brain activity and cognition: a connection from thermodynamics and information theory. *Frontiers in psychology*, 6:818, 2015.

[30] Efthymios Constantinides. Influencing the online consumer's behavior: the web experience. *Internet research*, 14(2):111–126, 2004.

[31] Marco Conti, Sajal K Das, Chatschik Bisdikian, Mohan Kumar, Lionel M Ni, Andrea Passarella, George Roussos, Gerhard Tröster, Gene Tsudik, and Franco Zambonelli. Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber–physical convergence. *Pervasive and Mobile Computing*, 8(1):2–21, 2012.

[32] Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008.

[33] Suzanne Curtin, Toben H Mintz, and Morten H Christiansen. Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, 96(3):233–262, 2005.

[34] Antonio Damasio. Descartes' error: Emotion, rationality and the human brain. *New York: Putnam*, 352, 1994.

[35] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274, 2016.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[37] Justin Diamond. " genlangs" and zipf's law: Do languages generated by chatgpt statistically look human? *arXiv preprint arXiv:2304.12191*, 2023.

[38] Michele T Diaz and Gregory McCarthy. A comparison of brain activity evoked by single content and function words: an fmri investigation of implicit word processing. *Brain research*, 1282:38–49, 2009.

[39] Kevin Dunbar and Debra Sussman. Toward a cognitive account of frontal lobe function: Simulating frontal lobe deficits in normal subjects. *ANNALS-NEW YORK ACADEMY OF SCIENCES*, 769:289–304, 1995.

[40] R. I. M. (Robin Ian MacDonald) Dunbar. *Grooming, gossip, and the evolution of language*. Harvard University Press, 1996.

[41] RI Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 9(10):178–190, 1998.

[42] Robin Dunbar. Theory of mind and the evolution of language. *Approaches to the Evolution of Language*, 1998.

[43] Robin I. M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):178–190, 1998.

[44] Robin IM Dunbar, Valerio Arnaboldi, Marco Conti, and Andrea Passarella. The structure of online social networks mirrors those in the offline world. *Social networks*, 43:39–47, 2015.

[45] John Duncan, Robert Ward, and Kimron Shapiro. Direct measurement of attentional dwell time in human vision. *Nature*, 369(6478):313–315, 1994.

[46] Arnaud D'Argembeau. On the role of the ventromedial prefrontal cortex in self-processing: the valuation hypothesis. *Frontiers in human neuroscience*, 7:372, 2013.

[47] Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[48] Michael J Emerson and Akira Miyake. The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, 48(1):148–168, 2003.

[49] Adrienne L Fairhall, Geoffrey D Lewen, William Bialek, and Robert R de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

[50] Ludovic Ferrand and Boris New. Semantic and associative priming in the mental lexicon. *Mental lexicon: Some words to talk about words*, pages 25–43, 2003.

[51] Paul C Fletcher, Francesca Happe, Uta Frith, Simon C Baker, Ray J Dolan, Richard SJ Frackowiak, and Chris D Frith. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, 57(2):109–128, 1995.

[52] Robert H Frank. *Passions within reason: The strategic role of the emotions.* WW Norton & Co, 1988.

[53] Flavius Frasincar, Jethro Borsje, and Leonard Levering. A semantic web-based approach for building personalized news services. *International Journal of E-Business Research (IJEBR)*, 5(3):35–53, 2009.

[54] Angela D Friederici, Bertram Opitz, and D Yves Von Cramon. Segregating semantic and syntactic aspects of processing in the human brain: an fmri investigation of different word types. *Cerebral cortex*, 10(7):698–705, 2000.

[55] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

[56] Helen L Gallagher, Francesca Happé, Nicola Brunswick, Paul C Fletcher, Uta Frith, and Chris D Frith. Reading the mind in cartoons and stories: an fmri study of 'theory of mind'in verbal and nonverbal tasks. *Neuropsychologia*, 38(1):11–21, 2000.

[57] Shaun Gallagher. Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences*, 4(1):14–21, 2000.

[58] Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407, 2019.

[59] Vinod Goel and Raymond J Dolan. The functional anatomy of humor: segregating cognitive and affective components. *Nature neuroscience*, 4(3):237–238, 2001.

[60] Daniel G Goldstein and Gerd Gigerenzer. Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75, 2002.

[61] Robert L Goldstone and Gary Lupyan. Discovering psychological principles by mining naturally occurring data sets. *Topics in cognitive science*, 8(3):548–568, 2016.

[62] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, 2011.

[63] Cleotilde Gonzalez. Training decisions from experience with decision making games. *Adaptive technologies for training and education*, pages 167–178, 2012.

[64] Cleotilde Gonzalez. 13 decision-making: A cognitive science perspective. *The Oxford handbook of cognitive science*, page 249, 2016.

[65] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[66] Richard L Gregory. *Seeing through illusions.* Oxford University Press, 2009.

[67] Thomas L Griffiths. Manifesto for a new (computational) cognitive revolution. *Cognition*, 135:21–23, 2015.

[68] Kalanit Grill-Spector, Richard Henson, and Alex Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23, 2006.

[69] Maarten Grootendorst. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics., 2020.

[70] Jan O Haerter, Bjørn Jamtveit, and Joachim Mathiesen. Communication dynamics in finite capacity social networks. *Physical review letters*, 109(16):168701, 2012.

[71] Linda Hermer and Elizabeth S Spelke. A geometric process for spatial reorientation in young children. *Nature*, 370(6484):57–59, 1994.

[72] Linda Hermer-Vazquez, Elizabeth S Spelke, and Alla S Katsnelson. Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive psychology*, 39(1):3–36, 1999.

[73] Ralph Hertwig and Peter M Todd. More is not always better: The benefits of cognitive limits. *Thinking: Psychological perspectives on reasoning, judgment and decision making*, pages 213–231, 2003.

[74] William E Hick. On the rate of gain of information. *Quarterly Journal of experimental psychology*, 4(1):11–26, 1952.

[75] Russell A Hill and Robin IM Dunbar. Social network size in humans. *Human nature*, 14(1):53–72, 2003.

[76] Christopher K Hsee. Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, 11(2):107–121, 1998.

[77] HJ Hsu, MH Christiansen, JB Tomblin, X Zhang, and RL Gómez. Statistical learning of nonadjacent dependencies in adolescents with and without language impairment. In *Poster presented at the 2006 symposium on research in child language disorders, Madison, WI*, 2006.

[78] Russell T Hurlburt. *Sampling inner experience in disturbed affect*. Springer Science & Business Media, 1993.

[79] Russell T Hurlburt. *Sampling normal and schizophrenic inner experience*. Springer Science & Business Media, 2012.

[80] Ray Hyman. Stimulus information as a determinant of reaction time. *Journal of experimental psychology*, 45(3):188, 1953.

[81] William James. The principles of. *Psychology*, 2:94, 1890.

[82] George F Jenks. The data model concept in statistical mapping. *International yearbook of cartography*, 7:186–190, 1967.

[83] George F Jenks. Optimal data classification for choropleth maps. *Department of Geographiy, University of Kansas Occasional Paper*, 1977.

[84] Nirmal Jonnalagedda and Susan Gauch. Personalized news recommendation using twitter. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 21–25. IEEE, 2013.

[85] Daniel Kahneman, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[86] Daniel Kahneman and Amos Tversky. On the study of statistical intuitions. Technical report, STANFORD UNIV CA DEPT OF PSYCHOLOGY, 1981.

[87] Ryota Kanai, Bahador Bahrami, R Roylance, and Geraint Rees. Online social network size is reflected in human brain structure. In *Proc. R. Soc. B*, volume 279, pages 1327–1334. The Royal Society, 2012.

[88] Sherif Karama, André Roch Lecours, Jean-Maxime Leroux, Pierre Bourgouin, Gilles Beaudoin, Sven Joubert, and Mario Beauregard. Areas of brain activation in males and females during viewing of erotic film excerpts. *Human brain mapping*, 16(1):1–13, 2002.

[89] Yaakov Kareev, Iris Lieberman, and Miri Lev. Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, 126(3):278, 1997.

[90] Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. The discrimination of visual number. *The American journal of psychology*, 62(4):498–525, 1949.

[91] Veronika Krieghoff, Florian Waszak, Wolfgang Prinz, and Marcel Brass. Neural and behavioral correlates of intentional actions. *Neuropsychologia*, 49(5):767–776, 2011.

[92] Patricia K Kuhl. The special-mechanisms debate in speech research: Categorization tests on animals and infants. 1987.

[93] Pierre-Simon Laplace. *A philosophical essay on probabilities*. Courier Corporation, 2012.

[94] Shu-Yueh Lee, Sara Steffes Hansen, and Jin Kyun Lee. What makes us click "like" on facebook? examining psychological, technological, and motivational factors on virtual endorsement. *Computer Communications*, 73:332–341, 2016.

[95] Peter Lennie. The cost of cortical computation. *Current biology*, 13(6):493–497, 2003.

[96] Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38, 1999.

[97] William B Levy and Victoria G Calvert. Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number. *Proceedings of the National Academy of Sciences*, 118(18):e2008173118, 2021.

[98] Penelope A Lewis, Roozbeh Rezaie, Rachel Brown, Neil Roberts, and Robin IM Dunbar. Ventromedial prefrontal volume predicts understanding of others and social network size. *Neuroimage*, 57(4):1624–1629, 2011.

[99] Hans Liljenström. Intention and attention in consciousness dynamics and evolution. *J. Cosmology*, 14:4848–4858, 2011.

[100] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[101] Theodor Lipps. Die aufgabe der erkenntnistheorie und die wundt'sche logik i. *Philosophische Monatshefte*, 16:529–539, 1880.

[102] Miriam A Locher and Sage L Graham. *Interpersonal pragmatics*, volume 6. Walter de Gruyter, 2010.

[103] Lola L Lopes. The rhetoric of irrationality. *Theory & Psychology*, 1(1):65–82, 1991.

[104] Gary Lupyan. Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic bulletin & review*, 16:711–718, 2009.

[105] P. Mac Carron, K. Kaski, and R. Dunbar. Calling dunbar's numbers. *Social Networks*, 47:151–155, 2016.

[106] Paul D MacLean. On the origin and progressive evolution of the triune brain. In *Primate brain evolution*, pages 291–316. Springer, 1982.

[107] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[108] René Marois and Jason Ivanoff. Capacity limits of information processing in the brain. *Trends in cognitive sciences*, 9(6):296–305, 2005.

[109] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov 2017.

[110] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[111] Sebastian Mežnar, Matej Bevec, Nada Lavrač, and Blaž Škrlj. Link analysis meets ontologies: Are embeddings the answer? *arXiv preprint arXiv:2111.11710*, 2021.

[112] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[113] George A Miller. The cognitive revolution: a historical perspective. *Trends in cognitive sciences*, 7(3):141–144, 2003.

[114] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3:1950, 2013.

[115] Giovanna Miritello, Esteban Moro, Rubén Lara, Rocío Martínez-López, John Belchamber, Sam G.B. Roberts, and Robin I.M. Dunbar. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, jan 2013.

[116] IC Mogotsi. 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.

[117] Jorge Moll, Ricardo de Oliveira-Souza, Ivanei E Bramati, and Jordan Grafman. Functional networks in emotional moral and normoral social judgments. In *Social Neuroscience*, pages 63–72. Psychology Press, 2013.

[118] Walle JH Nauta. The problem of the frontal lobe: A reinterpretation. *Principles, Practices, and Positions in Neuropsychiatric Research*, pages 167–187, 1972.

[119] John G Neuhoff. An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, 13(2):87–110, 2001.

[120] Albert Newen. What are cognitive processes? an example-based approach. *Synthese*, 194(11):4251–4268, 2017.

[121] Elissa L Newport and Richard N Aslin. Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2):127–162, 2004.

[122] Kevin N Ochsner and James J Gross. The cognitive control of emotion. *Trends in cognitive sciences*, 9(5):242–249, 2005.

[123] Luca Onnis, Padraic Monaghan, Korin Richmond, and Nick Chater. Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, 53(2):225–237, 2005.

[124] Takahiro Osada, Akitoshi Ogawa, Akimitsu Suda, Koji Nakajima, Masaki Tanaka, Satoshi Oka, Koji Kamagata, Shigeki Aoki, Yasushi Oshima, Sakae Tanaka, et al. Parallel cognitive processing streams in human prefrontal cortex: Parsing areal-level brain network for response inhibition. *Cell Reports*, 36(12), 2021.

[125] Ferdinand Osterreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.

[126] A Ross Otto and Nathaniel D Daw. The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, 123:92–105, 2019.

[127] Richard E Passingham. *The human primate*. Wh Freeman, 1982.

[128] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[129] Jiaxin Pei and David Jurgens. Quantifying intimacy in language. *arXiv preprint arXiv:2011.03020*, 2020.

[130] Marcela Peña, Luca L Bonatti, Marina Nespor, and Jacques Mehler. Signal-driven computations in speech processing. *Science*, 298(5593):604–607, 2002.

[131] Charles A Perfetti, Edward W Wlotko, and Lesley A Hart. Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1281, 2005.

[132] Luiz Pessoa. On the relationship between emotion and cognition. *Nature reviews neuroscience*, 9(2):148–158, 2008.

[133] Karl Magnus Petersson, Christian Forkstam, and Martin Ingvar. Artificial syntactic violations activate broca's region. *Cognitive science*, 28(3):383–407, 2004.

[134] Steven Pinker. *The language instinct: How the mind creates language*. Penguin UK, 2003.

[135] Irwin Pollack. The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, 24(6):745–749, 1952.

[136] Joanne Powell, Penelope A Lewis, Neil Roberts, Marta García-Fiñana, and Robin IM Dunbar. Orbital prefrontal cortex volume predicts social network size: an imaging study of individual differences in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1736):2157–2162, 2012.

[137] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[138] Qingqing Qu, Qingfang Zhang, and Markus F Damian. 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.

[139] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010.

[140] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

[141] Charles Ringer, Mihalis Nicolaou, and James Walker. Twitchchat: A dataset for exploring livestream chat. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pages 259–265, 2020.

[142] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[143] Jari Saramäki, Elizabeth A Leicht, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, and Robin IM Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[144] Elad Schneidman, Idan Segev, and Naftali Tishby. Information capacity and robustness of stochastic neuron models. *Advances in neural information processing systems*, 12, 1999.

[145] Eric Schulz, Rahul Bhui, Bradley C Love, Bastien Brier, Michael T Todd, and Samuel J Gershman. Exploration in the wild. *BioRxiv*, page 492058, 2018.

[146] Utlu I. Senel L. K., Koc A. Yucesoy V., and Cukur T. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[147] Anuj K Shah and Daniel M Oppenheimer. Heuristics made easy: An effort-reduction framework. *Psychological bulletin*, 134(2):207, 2008.

[148] Michael Siegal and Rosemary Varley. Neural systems involved in'theory of mind'. *Nature Reviews Neuroscience*, 3(6):463–471, 2002.

[149] Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956.

[150] Herbert A. Simon. Invariants of human behavior. *Annual Review of Psychology*, 41(1):1–20, 1990. PMID: 18331187.

[151] Herhert A Simon. The recognition heuristic how ignorance makes us smart. *Simple heuristics that make us smart*, page 37, 1999.

[152] Burrhus Frederic Skinner. *Science and human behavior*. Number 92904. Simon and Schuster, 1965.

[153] Louis Sokoloff. Local cerebral energy metabolism: its relationships to local functional activity and blood flow. *Cerebral vascular smooth muscle and its control*, page 171, 2009.

[154] Oleg Solopchuk, Andrea Alamia, Etienne Olivier, and Alexandre Zénon. Chunking improves symbolic sequence processing and relies on working memory gating mechanisms. *Learning & Memory*, 23(3):108–112, 2016.

[155] Dan Sperber. The modularity of thought and the epidemiology of representations. *Mapping the mind: Domain specificity in cognition and culture*, 39:67, 1994.

[156] Olaf Sporns and Richard F Betzel. Modular brain networks. *Annual review of psychology*, 67:613–640, 2016.

[157] Jon Sprouse. A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, 43(1):155–167, 2011.

[158] Darryl Stewart, Philip Hanna, F Smith, et al. Zipf and type-token rules for the english, spanish, irish and latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics*, 1:1–12, 2006.

[159] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. Crowdsourcing samples in cognitive science. *Trends in cognitive sciences*, 21(10):736–748, 2017.

[160] Valerie E Stone, Simon Baron-Cohen, and Robert T Knight. Frontal lobe contributions to theory of mind. *Journal of cognitive neuroscience*, 10(5):640–656, 1998.

[161] Donald T Stuss, Gordon G Gallup Jr, and Michael P Alexander. The frontal lobes are necessary fortheory of mind'. *Brain*, 124(2):279–286, 2001.

[162] Keerthana Muthu Subash, Lakshmi Prasanna Kumar, Sri Lakshmi Vadlamani, Preetha Chatterjee, and Olga Baysal. Disco: A dataset of discord chat conversations for software engineering research. In *Proceedings of the 19th International Conference on Mining Software Repositories*, pages 227–231, 2022.

[163] Alistair Sutcliffe, Robin Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology*, 103(2):149–168, 2012.

[164] Alistair G Sutcliffe, Di Wang, and Robin IM Dunbar. Modelling the role of trust in social relationships. *ACM Transactions on Internet Technology (TOIT)*, 15(4):16, 2015.

[165] Deborah Talmi, Cheryl L Grady, Yonatan Goshen-Gottstein, and Morris Moscovitch. Neuroimaging the serial position curve: A test of single-store versus dual-store models. *Psychological science*, 16(9):716–723, 2005.

[166] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*, 2023.

[167] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[168] Peter M Todd and Gerd Gigerenzer. Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, 23(5):727–741, 2000.

[169] J Bruce Tomblin, Elina Mainela-Arnold, and Xuyang Zhang. Procedural learning in adolescents with and without specific language impairment. *Language Learning and Development*, 3(4):269–293, 2007.

[170] John Tooby, Leda Cosmides, and Jerome Barkow. The adapted mind. *Evolutionary psychology and the generation of culture. New York: Oxford University Press*, 1992.

[171] Mustafa Toprak, Chiara Boldrini, Andrea Passarella, and Marco Conti. Harnessing the power of ego network layers for link prediction in online social networks. *IEEE Transactions on Computational Social Systems*, 2022.

[172] Ross Tucker and Timothy D Noakes. The physiological regulation of pacing strategy during exercise: a critical review. *British journal of sports medicine*, 43(6):e1–e1, 2009.

[173] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[174] Onur Varol, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Feature engineering for social bot detection. In *Feature engineering for machine learning and data analytics*, pages 311–334. CRC Press, 2018.

[175] Petra Vetter and Albert Newen. Varieties of cognitive penetration in visual perception. *Consciousness and cognition*, 27:62–75, 2014.

[176] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[177] A Wald. A method of estimating plane vulnerability based on damage of survivors, crc 432, july 1980. *Center for Naval Analyses*, 1980.

[178] Szu-Han Wang and Richard GM Morris. Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation. *Annual review of psychology*, 61:49–79, 2010.

[179] Yingxu Wang, Dong Liu, and Ying Wang. Discovering the capacity of human memory. *Brain and Mind*, 4:189–198, 2003.

[180] Peter C Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140, 1960.

[181] Kirsten Weber, Morten H Christiansen, Karl Magnus Petersson, Peter Indefrey, and Peter Hagoort. fmri syntactic and lexical repetition effects reveal the initial stages of learning a new language. *Journal of Neuroscience*, 36(26):6872–6880, 2016.

[182] Alexandre Zenon, Oleg Solopchuk, and Giovanni Pezzulo. An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123:5–18, 2019.

[183] Xiaohan Zhao, Alessandra Sala, Christo Wilson, Xiao Wang, Sabrina Gaito, Haitao Zheng, and Ben Y Zhao. Multi-scale dynamics in a massive online social network. In *Proceedings of the 2012 Internet Measurement Conference*, pages 171–184. ACM, 2012.

[184] W-X Zhou, D Sornette, R a Hill, and R I M Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings. Biological sciences / The Royal Society*, 272(1561):439–444, 2005.

[185] W.-X. Zhou, D. Sornette, R. A. Hill, and R. I. M. Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1561):439–444, 2005.

[186] Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*, 2021.

[187] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 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*, pages 19–27, 2015.

[188] George Kingsley Zipf. Human behavior and the principle of least effort. 1949.