TUDelft

Delft University of Technology

Topological properties and organizing principles of semantic networks

Budel, G.J.A.; Jin, Y.; Van Mieghem, P.F.A.; Kitsak, M.A.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

scientific reports

Check for updates

OPEN

Topological properties and organizing principles of semantic networks

Gabriel Budel[1,2], Ying Jin[1,2], Piet Van Mieghem[1] & Maksim Kitsak[1✉]

Interpreting natural language is an increasingly important task in computer algorithms due to the growing availability of unstructured textual data. Natural Language Processing (NLP) applications rely on semantic networks for structured knowledge representation. The fundamental properties of semantic networks must be taken into account when designing NLP algorithms, yet they remain to be structurally investigated. We study the properties of semantic networks from ConceptNet, defined by 7 semantic relations from 11 different languages. We find that semantic networks have universal basic properties: they are sparse, highly clustered, and many exhibit power-law degree distributions. Our findings show that the majority of the considered networks are scale-free. Some networks exhibit language-specific properties determined by grammatical rules, for example networks from highly inflected languages, such as e.g. Latin, German, French and Spanish, show peaks in the degree distribution that deviate from a power law. We find that depending on the semantic relation type and the language, the link formation in semantic networks is guided by different principles. In some networks the connections are similarity-based, while in others the connections are more complementarity-based. Finally, we demonstrate how knowledge of similarity and complementarity in semantic networks can improve NLP algorithms in missing link inference.

Due to the explosive increase in the availability of digital content, the demand for computers to efficiently handle textual data has never been greater. Large amounts of data and improved computing power have enabled a vast amount of research on Natural Language Processing (NLP). The goal of NLP is to allow computer programs to interpret and process unstructured text. In computers, text is represented as a string, while in reality, human language is much richer than just a string. People relate text to various concepts based on previously acquired knowledge. To effectively interpret the meaning of a text, a computer must have access to a considerable knowledge base related to the domain of the topic[1].

Semantic networks can represent human knowledge in computers, as first proposed by Quillian in the 1960s[2,3]. 'Semantic' means 'relating to meaning in language or logic' and a semantic network is a graph representation of structured knowledge. Such networks are composed of nodes, which represent concepts (*e.g.*, words or phrases), and links, which represent semantic relations between the nodes[4,5]. The links are tuples of the format *(source, semantic relation, destination)* that encode knowledge. For example, the information that a car has wheels is represented as *(car, has, wheels)*. Figure 1 shows a toy example of a semantic network as the subgraph with the neighborhood around the node *car*.

The past two decades have witnessed a rise in the importance of NLP applications[6–8]. For instance, Google introduced Google Knowledge Graph to enhance their search engine results[9]. A knowledge graph is a specific type of semantic network, in which the relation types are more explicit[10,11]. Voice assistants and digital intelligence services, such as Apple Siri[12] and IBM Watson[13], use semantic networks as a knowledge base for retrieving information[14,15]. As a result, machines can process information in raw text, comprehend unstructured user input, and achieve the goal of communicating with users, all up to a certain extent. Recently, OpenAI made a great leap forward in user-computer interaction with InstructGPT, better known to the general public as ChatGPT[16].

Language is a complex system with diverse grammatical rules. To grasp the meaning of a sentence, humans leverage their natural understanding of language and concepts in contexts. Language is still poorly understood from a computational perspective and, hence, it is difficult for computers to utilize similar strategies. Namely, machines operate under unambiguous instructions that are strictly predefined and structured by humans. Though we can argue that human languages are structured by grammar, these grammatical rules often prove

[1]Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands. [2]These authors contributed equally: Gabriel Budel and Ying Jin. ✉email: M.A.Kitsak@tudelft.nl
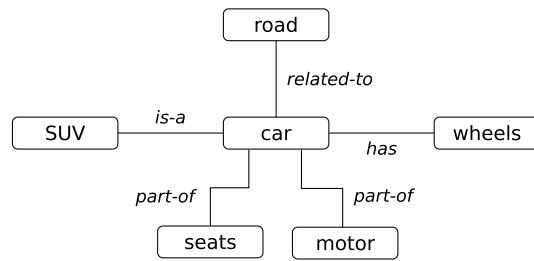
Figure 1. Toy example of a semantic network with six concepts and five semantic relations of four different types.

to be ambiguous[17]. After all, in computer languages, there are no synonyms, namesakes, or tones that can lead to misinterpretation[18]. Computers rely on external tools to enable the processing of the structure and meaning of texts.

In this paper, we conduct systematic analyses of the topological properties of semantic networks. Our work is motivated by the following purposes:

- Understand fundamental formation principles of semantic networks.
    In many social networks connections between nodes are driven by similarity[19–22]. The more similar two nodes are in terms of common neighbors, the more likely they are connected. Thanks to the intensive study of similarity-based networks, many successful tools of data analysis and machine learning were developed, such as link prediction[23] and community detection[24]. These tools may not work well for semantic networks, because words in a sentence do not necessarily pair together because of similarity. Sometimes, two words are used in conjunction because they have complementary features. Therefore, we study the principles that drive the formation of links in semantic networks.
- Document language-specific features.
    Languages vary greatly between cultures and across time[25]. Two languages that originate from two different language families can differ in many types of features since they are based on different rules. It is natural to conjecture that there exist diverse structures in semantic networks for different languages.
- Better inform NLP methods.
    Although there have been numerous real-world NLP applications across various domains, existing NLP technologies still have limitations[26]. For example, processing texts from a language where single words or phrases can convey more than one meaning is difficult for computers[27,28]. Existing, successful algorithms built on top of semantic networks are usually domain-specific, and designing algorithms for broader applications remains an open problem. To design better language models that can handle challenges such as language ambiguity, we first need to gain a better understanding of the topological properties of semantic networks.

Previous studies on semantic networks focused on a few basic properties and relied on multiple datasets with mixed semantic relations, which we discuss in detail in the 'Related work' section. Therefore, it is difficult to compare the results within one study and between two different studies. To our knowledge, there has been no systematic and comprehensive analysis of the topological properties of semantic networks at the semantic relation level.

To sum up, the main objective of this paper is to understand the structure of semantic networks. Specifically, we first study the general topological properties of semantic networks from a single language with distinct semantic relation types. Second, we compare semantic networks with the same relation type between different languages to find language-specific patterns. In addition, we investigate the roles of similarity and complementarity in the link formation principles in semantic networks.

The main contributions of this paper include:

1. We study the topological properties of seven English semantic networks, each network defined by a different semantic relation (*e.g.*, 'Is-A' and 'Has-A'). We show that all networks possess high sparsity and many possess a power-law degree distribution. In addition, we find that most networks have a high average clustering coefficient, while some networks show the opposite.
2. We extend the study of the topological properties of semantic networks to ten other languages. We find non-trivial structural patterns in networks from languages that have many grammatical inflections. Due to the natural structure of grammar in these languages, words have many distinct inflected forms, which leads to peaks in the density of the degree distribution and results in deviations from a power law. We find this feature not only in inflecting languages but also in Finnish, which is classified as agglutinating.
3. We study the organizing principles of 50 semantic networks defined by different semantic relations in different languages. We quantify the structural similarity and complementarity of semantic networks by counting the relative densities of triangles and quadrangles in the graphs, following a recent work by Talaga and Nowak[29]. Hereby, we show to what extent these networks are similarity- or complementarity-based. We find that the connection principles in semantic networks are mostly related to the type of semantic relation, not the language origin.

This paper is organized in the following manner: In the 'Related work' section, we provide a brief overview of the previous work on the properties of semantic networks. In the 'General properties of semantic networks' section, we study the general topological properties of seven English semantic networks. In the 'Language-specific properties' section, we compare the properties of semantic networks between 11 different languages. The section 'Similarity and complementarity in semantic networks' deals with the fundamental connection principles in semantic networks. We measure and compare the structural similarity and complementarity in the networks in this study and we discuss the patterns that arise. Finally, we summarize our conclusions and findings and give recommendations for future research in the 'Discussion' section.

Related work

Due to the growing interest in semantic networks, related studies were carried out in a wide range of different fields. Based on our scope, we focus on two main aspects in each work: the topological properties that were analyzed in the study and the dataset that was used in the analysis (i) and the universal and language-specific patterns which were found and discussed (ii).

The majority of semantic networks literature is centered around three link types: co-occurrence, association and semantic relation. In a co-occurrence network, sets of words that co-occur in a phrase, sentence or text form a link. For association networks, participants in a cognitive-linguistic experiment are given a word and asked to give the first word that they think of. There are several association datasets, one example is the University of South Florida Free Association Norms[30]. Semantic relations are relations defined by professionals like lexicographers, typical examples are synonym, antonym, hypernym and homonymy. The specific instances of the semantic relations are also defined by the lexicographers or extracted computationally from text corpora.

In 2001, Ferrer-i-Cancho and Sole[31] studied undirected co-occurrence graphs constructed from the British National Corpus dataset[32]. They measured the average distance between two words and observed the small-world property, which was found in many real-world networks[33]. Motter et al.[34] analyzed an undirected conceptual network constructed from an English Thesaurus dictionary[35]. They focused on three properties: sparsity (small average degree), average shortest path length and clustering. That same year, Sigman and Cecchi[36] studied undirected lexical networks extracted from the noun subset of WordNet[37], where the nodes are sets of noun synonyms. They grouped networks by three semantic relations: antonymy, hypernymy and meronymy. A detailed analysis of characteristic length (the median minimal distance between pairs of nodes), degree distributions and clustering of these networks were provided. Semantic networks were also found to possess the small-world property of sparse connectivity, short average path length, and strong local clustering[34,36].

Later, Steyvers and Tenenbaum[38] performed statistical analysis of 3 kinds of semantic networks: word associations[30], WordNet and Roget's Thesaurus[39]. Apart from the above-mentioned network properties, they also considered network connectedness and diameter. They pointed out that the small-world property may originate from the scale-free organization of the network, which exists in a variety of real-world systems[40-42].

As for patterns across different languages, Ferrer-i-Cancho et al.[43] built syntactic dependency networks from corpora (collections of sentences) for three European languages: Czech, German and Romanian. They showed that networks from different languages have many non-trivial topological properties in common, such as the small-world property, a power-law degree distribution and disassortative mixing[44].

Existing studies have identified some general network properties in semantic networks such as the small-world property and power-law degree distributions. However, the datasets used in these studies are often different, sometimes even within the same study, rendering direct comparison of results difficult. Some used associative networks generated from experiments and some studied thesauri that were manually created by linguists. In addition, most of the research performed consists of coarse-grained statistical analyses. Specifically, different semantic relations were sometimes treated as identical and the subset of included nodes was often limited (e.g., only words and no phrases or only nouns). Further, there are only very few studies on semantic networks from languages other than English.

Therefore, our analyses focus on semantic networks with different semantic relations (link types) from a single dataset. We consider networks defined by a specific link type, make these networks *undirected* and *unweighted* and compare the structural properties between networks with different link types. In addition, we apply similar analyses to semantic networks with the same link type across different languages. Furthermore, we investigate the roles that similarity and complementarity play in the formation of links in semantic networks.

General properties of semantic networks

To gain an understanding of the structure of semantic networks, we first study their general topological properties. We introduce the main characteristics of the dataset that we use throughout this study, ConceptNet[45], in the section 'Data' in SI. Next, we list the semantic relations that define the networks in this study in Supplementary Table S1. The overview of the semantic networks is given in Supplementary Table S2. In this section, we compute various topological properties of these networks related to connectedness, degree, assortative mixing and clustering. We summarize the overall descriptive statistics of the semantic networks in Supplementary Table S3.

Connectedness. We measure the connectedness of a network by the size of the largest connected component and the size distribution of all connected components. The complete component size distributions of the English semantic networks are shown in Supplementary Figure S2. Supplementary Table S4 lists the sizes of the largest connected components (LCCs) in absolute numbers as well as relative to the network size. The same statistics are computed after degree-preserving random rewiring of the links for comparison[44]. The purpose of random rewiring is to estimate the value of a graph metric that could be expected by chance, solely based on the node degrees (see SI for details on the rewiring process).

Based on the percentages of nodes in the LCC, all seven semantic networks are not fully connected. The networks 'Is-A', 'Related-To' and 'Union' are almost fully connected given that their LCCs contain over 90% of nodes. Networks 'Has-A', 'Part-Of', 'Antonym' and 'Synonym' are largely disconnected, with the percentages of nodes in their LCCs ranging from 22% to 65%. Most of the rewired networks are more connected than the corresponding original networks, especially networks 'Antonym' and 'Synonym'. In other words, the majority of our semantic networks are less connected than what could be expected by chance. For networks 'Related-To' and 'Union', the percentage of nodes in the LCC remains almost unchanged, while the 'Is-A' network is more connected than expected.

Degree distribution.   Figure 2 shows that the densities $Pr[D = k]$ of the degree distributions of our seven English semantic networks all appear to approximately follow power laws in the tail visually. A more rigorous framework for assessing power laws was proposed by Voitalov et al.[46], who consider networks to have a power-law degree distribution if $Pr[D = k] = \ell(k)k^{-\gamma}$ for a slowly varying function $\ell(k)$, see the section 'Consistent power-law exponent estimators' in SI. Figure 2 includes the estimates $\hat{\gamma}$ based on the slopes of the densities $Pr[D = k]$ on a log-log scale, along with the three consistent estimators from the framework of Voitalov et al.[46,47]. According to these estimators, the degree sequences of 5 out of the 7 networks are power-law. The degree sequences of the 'Synonym' and 'Antonym' networks are *hardly power-law* because at least one of the $\hat{\gamma} > 5$ and therefore the estimated exponents are not listed.

For most networks, the estimated exponent $\hat{\gamma}$ lies between 2 and 3. Therefore, most semantic networks are scale-free, except for the 'Synonym' and 'Antonym' networks. In the literature, semantic networks were also found to be highly heterogeneous[38,48]. Moreover, the word frequencies in several modern languages were found to follow power laws[49]. In the section 'Language-specific properties', we will see that the 'Synonym' and 'Antonym' networks in most considered languages are hardly power-law or not power-law networks.

The heterogeneity in the degree distribution seems natural for networks such as the 'Is-A' network: there are many specific or unique words with a small degree that connect to only a few other words, while there are also a few general words that connect to almost anything, resulting in a large degree. Examples of general words with a large degree are 'plant' and 'person', while specific words like 'neotectonic' and 'cofinance' have a small degree. Our results show that many semantic networks have power-law degree distributions, like many other types of real-world networks[50–52].

Degree assortativity.   A number of measures have been established to quantify degree assortativity, such as the degree correlation coefficient ρ_D and the Average Nearest Neighbor Degree (ANND)[44]. Figure 3 shows the average nearest neighbor degree as a function of the degree k for four selected networks and their values after
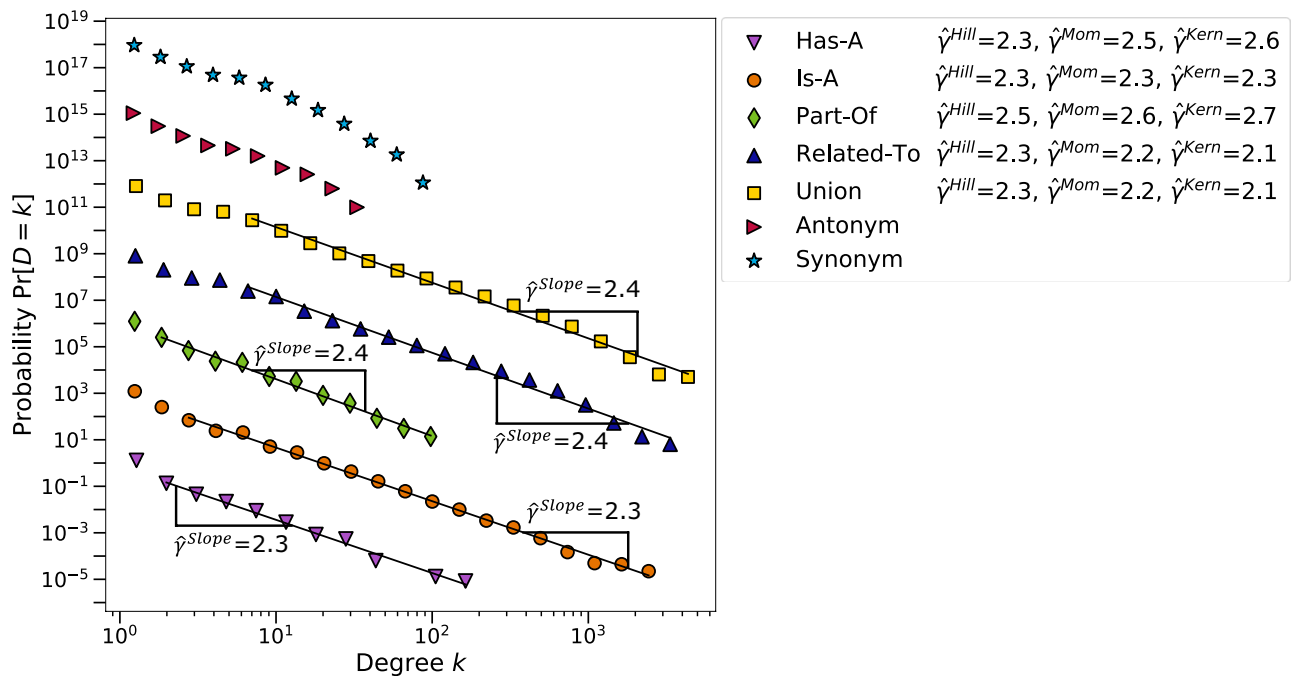
Figure 2. Degree distribution densities $Pr[D = k]$ of the LCCs of the seven English semantic networks. The data is scaled by powers of 1000 to better visualize the power law in each density. The corresponding estimated power-law exponents $\hat{\gamma}$ are shown if there is a power law, $Pr[D = k] \approx \ell(k)k^{-\gamma}$. The degree sequences of the networks 'Antonym' and 'Synonym' were estimated to be *hardly power-law* because at least one of the $\hat{\gamma} > 5$. The data are logarithmically binned to suppress noise in the tails of the distributions, see the section 'Power-law degree distributions' in SI for details on how the power-law densities are processed and the power-law exponent estimation procedures.
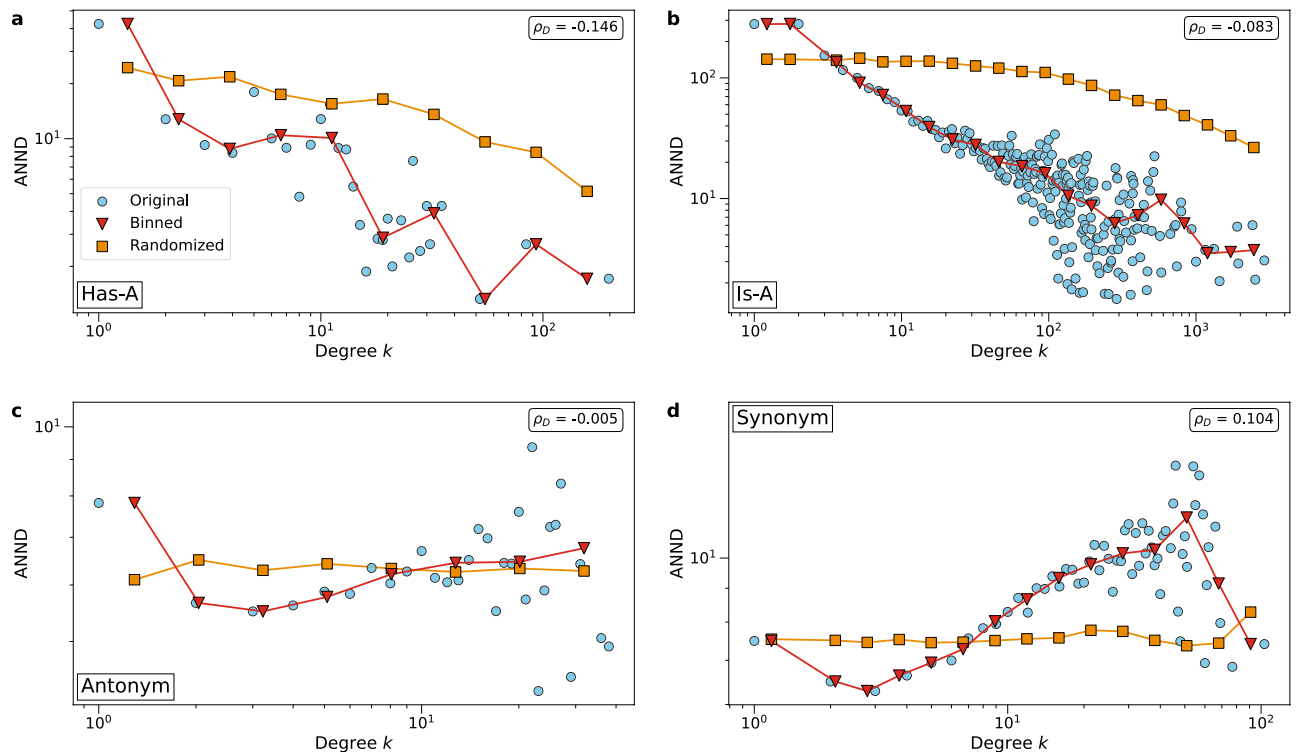
Figure 3. Average nearest neighbor degree (ANND) as a function of degree k and degree correlation coefficient ρ_D of four English semantic networks. (**a**) Network '*Has-A*', (**b**) Network '*Is-A*', (**c**) Network '*Antonym*', (**d**) Network '*Synonym*'. See SI for the results of all seven networks. Data points in circles are the average ANND of nodes with degree k in a network, triangles represent the data after logarithmic binning, and squares are the average ANND of nodes with degree k in the rewired network. Logarithmic binning is used to better visualize the data.

random rewiring as well as the degree correlation coefficient ρ_D. Refer to Supplementary Fig. S3 for the ANND plots of all networks. The randomized networks with preserved degree distribution have no degree-degree correlation. As a result, the function ANND does not vary with k. The randomized networks serve as a reference for the expected ANND values when the links are distributed at random.

We find that most semantic networks are disassortative as ANND is a decreasing function of the degree k and the degree correlation coefficient ρ_D is negative. These networks are 'Has-A', 'Part-Of', 'Is-A', 'Related-To' and 'Union'. In disassortative networks, nodes with larger degrees (general words) tend to connect to nodes with smaller degrees (less general words). This is not surprising. Indeed, if we use these relations in a sentence, then we often relate specific words to more general words. For example, we say 'horse racing is a sport', in which 'horse racing' is a very specific phrase while 'sport' is more general.

On the other hand, network 'Synonym' is assortative as the function ANND increases in the degree k. This indicates that large-degree nodes (general words) connect to nodes that have similar degree (words with the same generality). The same applies to network 'Antonym'. Although the degree correlation is not very pronounced and reflected by the small correlation coefficient $\rho_D = -0.005$, we still see a slight upward trend in the curve of ANND.

The function ANND of a rewired network is not degree-dependent anymore, shown by the orange curves in Figure 3. The curve is almost flat for 'Synonym' and 'Related-To'. At the larger degree k, the curve may drop slightly, as for large-degree nodes there are not enough nodes of equal degree to connect to.

Clustering coefficient.

In networks such as social networks, the neighbors of a node are likely to be connected as well, a phenomenon which is known as clustering[33,53]. If a person has a group of friends, there is a high chance that these friends also know each other. These networks are characterized by many triangular connections.

Figure 4 shows the average clustering coefficient $c_G(i)$ of nodes with degree $d_i = k$ of four English networks. Refer to Supplementary Fig. S4 for the clustering coefficients of all seven networks. All networks have small clustering coefficients in absolute terms, which, in combination with the small average degree $E[D]$, indicates a local tree-like structure. We find that the networks 'Part-Of', 'Antonym' and 'Synonym' have substantially larger clustering coefficients than their rewired counterparts: there are more triangles in these networks than expected by chance. On the other hand, the network 'Has-A' has lower clustering coefficients $c_G(i)$ than the randomized network, therefore it seems that the 'Has-A' network is organized in a different way than the other networks. As
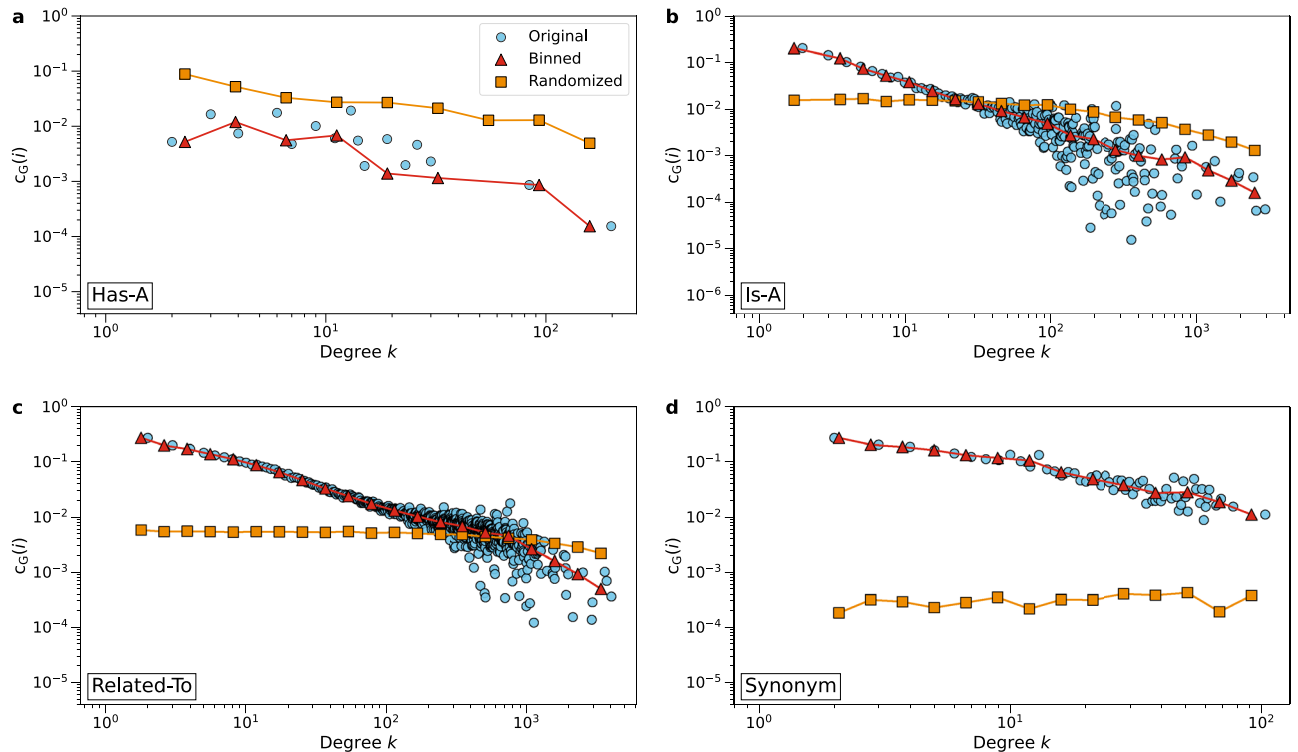
Figure 4. The average clustering coefficient $c_G(i)$ of nodes with degree $d_i = k$ of four English semantic networks. (**a**) Network *'Has-A'*, (**b**) Network *'Is-A'*, (**c**) Network *'Related-To'*, (**d**) Network *'Synonym'*. See the SI for the results of all seven networks. Data points in circles are the original average local clustering coefficients of nodes with degree $d_i = k$, triangles represent data after logarithmic binning, and squares show the average clustering coefficients of nodes with degree $d_i = k$ (logarithmically binned) in the randomized networks.

for the networks 'Is-A', 'Related-To' and 'Union', the clustering coefficients $c_G(i)$ are similar to their corresponding rewired networks.

In summary, we find that English semantic networks have power-law degree distributions and most are scale-free, which coincides with the results in previous studies[38,48]. Besides, semantic networks with different link types show different levels of degree assortativity and average local clustering. Most works in the literature have identified high clustering coefficients in semantic networks[34,36,38,48]. This encourages us to further investigate the organizing principles of these semantic networks, which we will discuss explicitly in the 'Similarity and complementarity in semantic networks' section.

Language-specific properties

Up to this point, we have only considered English semantic networks, while there are thousands of other languages in the world besides English. In this section, we consider semantic networks from 10 other languages contained within ConceptNet: French, Italian, German, Spanish, Russian, Portuguese, Dutch, Japanese, Finnish and Chinese. We group the in total 11 languages based on their language families and we again study the topological properties of 7 different semantic relations per language. Finally, we observe peculiarities in the degree distribution densities of the 'Related-To' networks in some languages, which we later explain by grammar.

Language classification. In linguistics, languages can be partitioned in multiple different ways. Mainly, there are two kinds of language classifications: genetic and typological.

The genetic classification assorts languages according to their level of diachronic relatedness, where languages are categorized into the same family if they evolved from the same root language.[54] An example is the Indo-European family, which includes the Germanic, Balto-Slavic and Italic languages[55].

One popular typological classification distinguishes isolating, agglutinating and inflecting languages. It groups languages in accordance with their morphological word formation styles. A morph or morpheme (the Greek word $\mu o \rho \varphi \acute{\eta}$ means 'outer shape, appearance' of which the English 'form' is derived) is the basic unit of a word, such as a stem or an affix[56]. For instance, the word 'undoubtedly' consists of three morphs: 'un-', 'doubted' and '-ly'. In an isolating language, like Mandarin Chinese, each word contains only a single morph[54]. In contrast, words from an agglutinating language can be divided into morphs with distinctive grammatical categories like tense, person and gender. In an inflecting language, however, there is no exact match between morphs and grammatical categories[54]. A word changes its form based on different grammar rules. Most Indo-European languages belong to the inflecting category.

Based on these two types of classifications, we have selected 11 languages to cover different language types, Supplementary Table S5. Typologically, Chinese is an isolating language, while Japanese and Finnish are agglutinating languages. The rest of the languages under consideration (8 out of 11) belong to the inflecting category. Genetically, French, Italian, Spanish and Portuguese belong to the Italic family, while English, German and Dutch are Germanic. Russian is a Balto-Slavic language, Japanese is Transeurasian, Chinese is Sino-Tibetan and Finnish belongs to the Uralic family. We mainly refer to the typological classification throughout our analyses.

Overview of semantic networks from eleven languages. For every language, we construct seven undirected semantic networks with the link types 'Has-A', 'Part-Of', 'Is-A', 'Related-To', 'Union', 'Antonym' and 'Synonym'. Due to missing data in ConceptNet, only three languages have the 'Has-A' relation. For these languages, the 'Union' network is the union of three link types: 'Part-Of', 'Is-A' and 'Related-To'. In this section, we provide an overview of the numbers of nodes N and numbers of links L of the semantic networks. Again, we restrict our study to the LCCs of these networks.

Regarding the numbers of nodes N, the networks 'Related-To' and 'Union' are generally the largest networks in a language, with the French 'Union' network being the absolute largest with $N = 1,296,622$, as denoted in Supplementary Table S6. Nevertheless, there are many small networks with size $N < 100$, particularly for the 'Part-Of' and 'Synonym' networks.

Similar to the English semantic networks, we observe that most networks with more than 100 nodes are sparse. All networks have an average degree between 1 and 6, which is small compared to the network size. Consider the Dutch 'Is-A' network for example, where a node has about 5 connections on average, which is only 2.45% of 191 nodes in the whole network. Supplementary Table S7 lists the average degree $E[D]$ of all our semantic networks.

Degree distribution. Many of the semantic networks in the 11 languages have degree distributions that are approximately power laws. We estimate the power-law exponents only for networks with size $N > 1000$ because we require a sufficient number of observations to estimate the power-law exponent γ. Supplementary Table S8 lists the estimated power-law exponents $\hat{\gamma}$ using the same 4 methods as in Fig. 2 for each semantic network. Refer to the section 'Power-law degree distributions' in SI for details on these estimation procedures.

We find that many networks have power laws in their degree distributions and many of those networks are scale-free ($2 < \hat{\gamma} < 3$). The Chinese 'Related-To' network even has a power-law exponent $\hat{\gamma} < 2$. The degree distributions of all 'Synonym' and 'Antonym' networks are hardly or not power laws, however. The likely reason for this is that nodes in these networks generally have smaller degrees than in other networks. As a result, the slope of the degree distribution is steeper and therefore not classified as a power law. This is not unexpected, as for a given word there are only a certain number of synonyms or antonyms and therefore there are not many nodes with high degrees. Another interesting finding is that the densities of the degree distributions of the 'Related-To' and 'Union' networks for French, Spanish, Portuguese and Finnish show notable deviations from a straight line in the log-log plot, which we discuss in-depth in the next section.

Language inflection. In some languages, the densities of the degree distributions of the 'Related-To' and 'Union' networks show deviations from a straight line on a log-log scale. An example is the Spanish 'Related-To' network in Fig. 5a, where we observe a peak in the tail of the distribution. To find the cause of the anomaly in the degree distribution, we inspect the words with a degree k located in the peak, referred to as *peak words*, and their neighbors. Supplementary Table S9 lists a few examples of the peak words, which are almost all verbs and have similar spellings. The links adjacent to these nodes with higher-than-expected degrees might be the result of grammatical inflections of the same root words since Spanish is a highly inflected language. We observe a similar anomaly in the degree distributions of French, Portuguese and Finnish 'Related-To' and 'Union' networks. In Supplementary Table S6 we saw that the network 'Union' is mostly composed of 'Related-To' in these four languages, therefore we restrict the analysis to the 'Related-To' networks.

Two common types of language inflection are conjugation, the inflection of verbs, and declension, the inflection of nouns. The past tense of the verb 'to sleep' is 'slept', an example of conjugation in English. The plural form of the noun 'man' is 'men', an example of declension. The languages Spanish, Portuguese and French are much richer in conjugations than English, while Finnish is rich in declensions.

Part-of-speech tags. In the ConceptNet dataset, only part of the nodes is part-of-speech (POS) tagged with one of four types: verb, noun, adjective and adverb. For French, Spanish and Portuguese, the percentage of verbs in the peak is larger than in the LCC, while for Finnish the percentage of nouns in the peak is larger than in the LCC, see Supplementary Table S10. Remarkably, 100% of the Portuguese peak words are verbs. Most neighbors of the peak words are verbs for Spanish (97%), Portuguese (99%) and French (87%), while most neighbors of the peak words are nouns for Finnish (90%), Supplementary Table S11. This strengthens our belief that the abnormal number of nodes with a certain degree k is related to language inflection in these four languages.

Merging of word inflections. To investigate whether the peaks in the degree distribution densities are indeed related to word inflections, we leverage the 'Form-Of' relation type in ConceptNet, which connects two words A and B if A is an inflected form of B, or B is the root word of A[57]. We merge each node and its neighbors from the 'Form-Of' network (its inflected forms) into a single node in the 'Related-To' network, as depicted in Supplementary Figure S5. Figure 5 shows the degree distribution densities of the 'Related-To' networks before and after node merging. The range of the anomalous peak in the density of the degree distribution is highlighted in yellow. In each panel, the number of grammatical variations m coincides with the center of the peak. As seen in
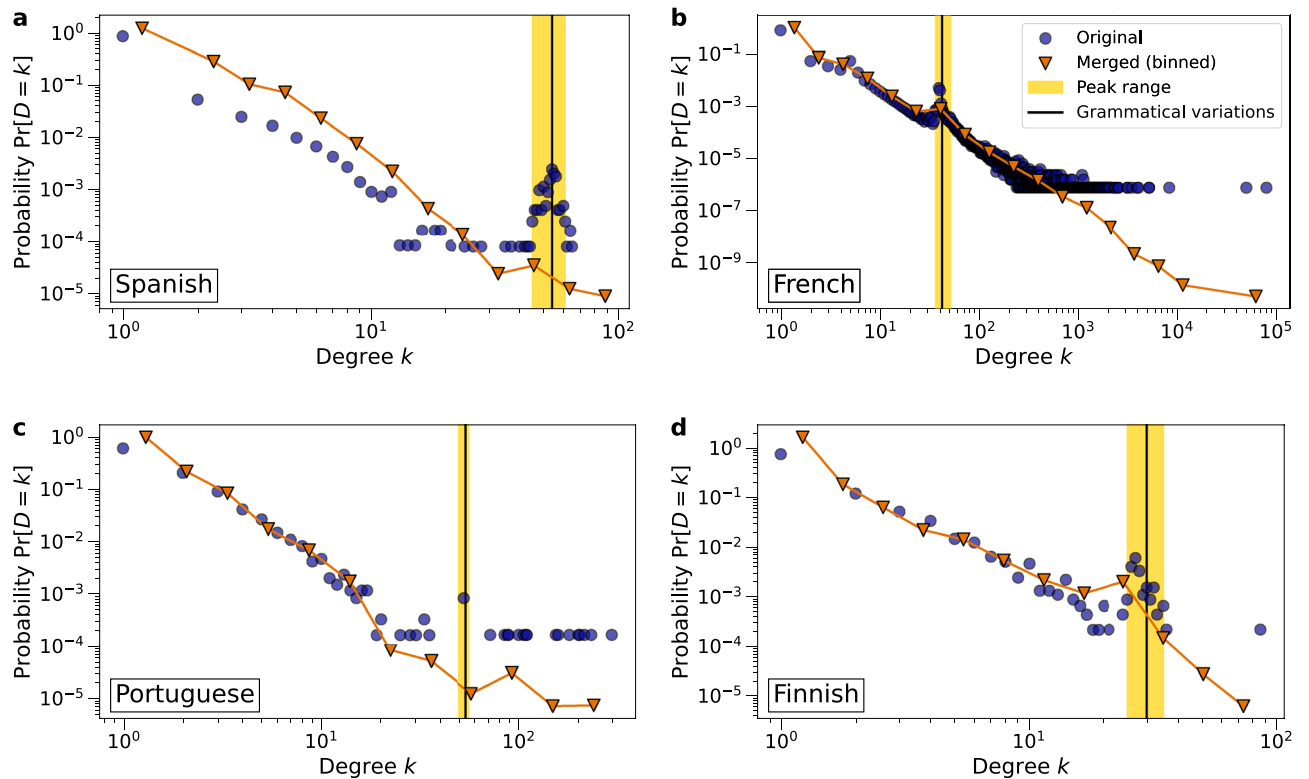
Figure 5. Densities $\Pr[D = k]$ of the degree distributions of the 'Related-To' networks before and after node merging of inflected forms in (**a**) Spanish, (**b**) French, (**c**) Portuguese and (**d**) Finnish. The logarithmically binned densities after node merging are shown in orange. The peaks are highlighted in yellow. The vertical black lines indicate the number of grammatical variations m for the relevant grammatical rule in the respective language. In each panel the number of grammatical variations m coincides with the center of the peak.

Fig. 5a, the peak completely disappears in the Spanish 'Related-To' after node merging, thus the peak is described entirely by connections due to word inflections. We also observe significant reductions in the heights of the peaks for Portuguese and Finnish 'Related-To' networks. However, for the French 'Related-To' network there is only a slight reduction in height after merging, which we believe is likely due to poor coverage in the French 'Form-Of' network with only 17% of words in the peak. In contrast, the Spanish 'Form-Of' network covers 97% of the Spanish peak words, while for Portuguese and Finnish approximately 50% of the peak words are covered, Supplementary Table S12.

The number of inflections. In a language, the number of distinct conjugations of regular verbs is determined by the number of different pronouns and the number of verb tenses, which are grammatical time references[58]. In Spanish, there are 6 pronouns and 9 simple verb tenses, resulting in at most $m = 6 \times 9 = 54$ distinct verb conjugations[59,60]. Table 1 exemplifies these 54 different conjugations for the verb 'amar', which means 'to love'. There are also irregular verbs that follow different, idiosyncratic grammatical rules, but the majority of the verbs in Spanish are classified as regular, like in most inflecting languages. The number of grammatical variations $m = 54$ coincides with the center of the peak in Figure 5a.

Like Spanish, Portuguese has $m = 54$ distinct conjugated verb forms[61]. In French, there are $m = 6 \times 7 = 42$ distinct verb conjugations[62]. In Finnish, there are in total 15 noun declensions or cases with distinct spelling, each having singular and plural forms, resulting in $m = 30$ different cases of a Finnish noun[63]. Supplementary Table S13 lists the number of grammatical variations m in French, Spanish, Portuguese and Finnish, along with the minimum and maximum degree k_{min} and k_{max} where the peak starts and ends. By Fig. 5 we confirm that the number of grammatical variations m coincides with or is close to the center of the peak.

In summary, we observe anomalies in the degree distributions of 'Related-To' networks from the inflecting languages Spanish, French and Portuguese and the agglutinating Finnish. Because of grammatical structures, root words in these languages share many links with their inflected forms, resulting in more nodes with a certain degree than expected. While Finnish is typologically classified as agglutinating, it still has many noun declensions, suggesting that the agglutinating and inflecting language categories are not mutually exclusive.

Tense	Pronoun					
	Yo	Tú	Él/Ella/Usted	Nosotros	Vosotros	Ellos/Ellas/Ustedes
	(I)	(You)	(He/She)	(We)	(You)	(They)
Present Indicative	amo	amas	ama	amamos	amáis	aman
Imperfect Indicative	amaba	amabas	amaba	amábamos	amabais	amaban
Preterite Indicative	amé	amaste	amó	amamos	amasteis	amaron
Future Indicative	amaré	amarás	amará	amaremos	amaréis	amarán
Conditional Indicative	amaría	amarías	amaría	amaríamos	amaríais	amarían
Present Subjunctive	ame	ames	ame	amemos	améis	amen
Imperfect Subjunctive 1	amara	amaras	amara	amáramos	amarais	amaran
Imperfect Subjunctive 2	amase	amases	amase	amásemos	amaseis	amasen
Future Subjunctive	amare	amares	amare	amáremos	amareis	amaren

Table 1. Verb conjugation table for the Spanish verb 'amar' (to love). The 6 pronouns and 9 verb tenses result in a maximum of 54 different conjugated forms.

Similarity and complementarity in semantic networks

Although we have identified several universal characteristics of semantic networks, we also observe notable differences in some of their properties. The clustering coefficient in some semantic networks, for instance, is greater than expected by chance, while in other semantic networks, *e.g.*, the English 'Has-A' network, the clustering coefficient is smaller than expected by chance.

We hypothesize that these semantic networks are organized according to different principles. It is commonly known that one such principle is similarity: all factors being equal, similar nodes are more likely to be connected. Similarity is believed to play a leading role in the formation of ties in social networks and lies at the heart of many network inference methods. At the same time, recent works indicate that many networks may be organized predominantly according to the complementarity principle, which dictates that interactions are preferentially formed between nodes with complementary properties. Complementarity has been argued to play a significant role in protein-protein interaction networks[64] and production networks[65]. In addition, a geometric complementarity framework for modeling and learning complementarity representations of real networks was recently formulated by Budel and Kitsak[66].

This section aims to assess the relative roles of complementarity and similarity mechanisms in different semantic networks. We utilize the method by Talaga and Nowak[29]. The method assesses the relative roles of the two principles by measuring the relative densities of triangular and quadrangle motifs in the networks. Intuitively, the transitivity of similarity - *A* similar to *B* and *B* similar to *C* implies *A* similar to *C* – results in a high density of triangles[20,67,68], Supplementary Figure S6a. The non-transitivity of complementarity, on the other hand, suppresses the appearance of triangles but enables the appearance of quadrangles in networks[64,69].

We measure and compare the density of triangles and quadrangles with the structural similarity and complementarity coefficients using the framework of Talaga and Nowak[29]. After computing the densities of triangles and quadrangles, the framework assesses their significance by comparing the densities to those of the configurational models built with matching degree distributions, see the SI for a summary. As a result of the assessment, the network of interest is quantified by two normalized structural coefficients corresponding to complementarity and similarity.

Figure 6 depicts the relative roles of complementarity and similarity in 50 semantic networks. We observe that semantic networks are clustered together according to semantic relation types and not their language families, indicating that specific types of semantic relations matter more for the organizing principles of a semantic network rather than its language.

Based on the calibrated complementarity and similarity values, we can categorize most semantic networks as (i) predominantly complementarity-based, (ii) predominantly similarity-based, and (iii) networks where both complementarity and similarity are substantially present.

We observe four clustering patterns in Figure 6.

- Cluster 1 (light blue): the 'Synonym' networks are characterized by stronger similarity than complementarity values. This observation is hardly surprising since 'Synonym' networks link words with similar meanings. Since similarity is transitive, the Synonym' networks contain significant numbers of connected node triples, leading to large clustering coefficients.
- Cluster 2 (red): the 'Antonym' networks, as observed in Fig. 6, belong to the upper triangle of the scatter plot plane, hence complementarity is more prevalent in these networks than similarity. This observation is as expected, as antonyms are word pairs with opposite meanings that complement each other. In our earlier work[66] we learned a geometric representation of the English 'Antonym' network demonstrating that antonyms indeed complement each other.

    More surprisingly, some antonym networks are characterized by substantial similarity values, implying the presence of triangle motifs. This is the case since there are instances of three or more words that have opposite meanings to all other words in the group. One example is the triple of words (*man*, *woman*, *girl*). Each pair of words in the triple is opposite in meaning along a certain dimension, here either gender or age.
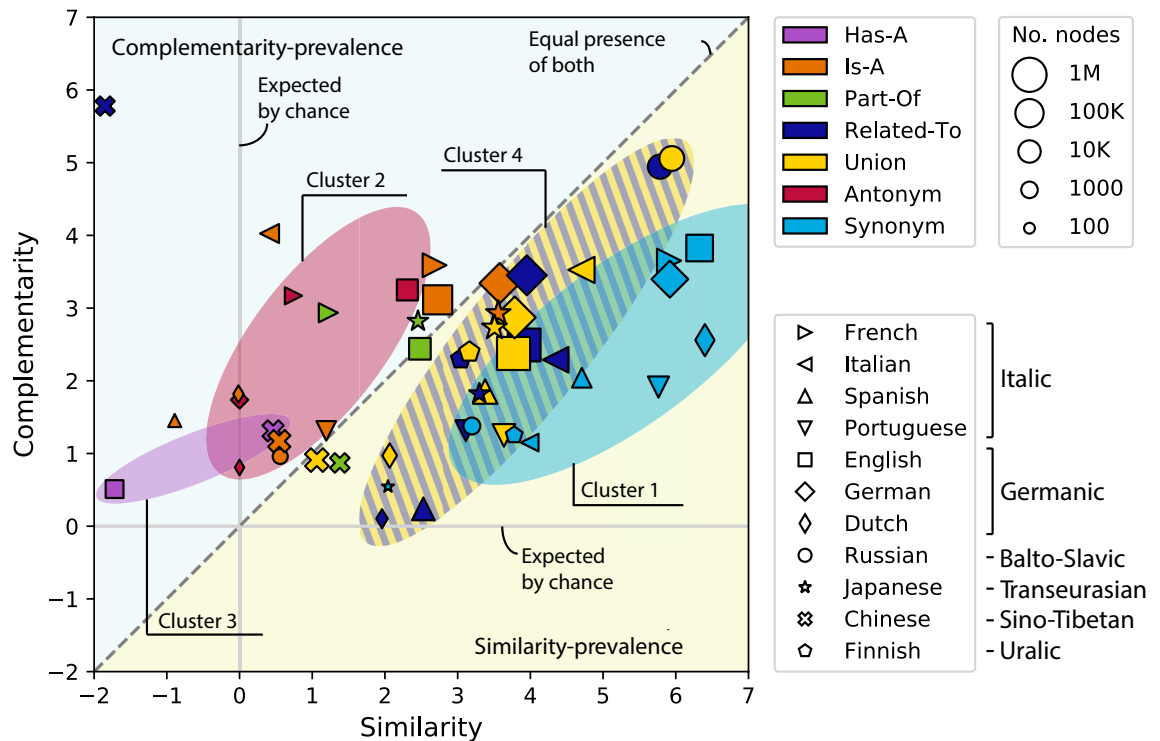
Figure 6. Calibrated average structural coefficients of the LCCs of the 50 semantic networks from 11 languages. Languages that belong to the same family are marked with similar shapes. Triangles represent Italic, quadrilaterals represent Germanic, circles represent Balto-Slavic, a star represents Transeurasian, a cross represents Sino-Tibetan and a pentagon represents Uralic. The marker size scales logarithmically with the number of nodes N in the network and is further adjusted for visibility. The grey lines at $x = 0$ and $y = 0$ indicate the expected coefficients based on the configuration model (see SI). The dashed line at $y = x$ indicates that the structural similarity and complementarity coefficients are equal. Networks in the upper left area (shaded in blue) are more complementarity-based, while networks in the lower right area (shaded in yellow) are more similarity-based. We highlight four clusters of networks using different colors.

- Cluster 3 (purple): the 'Has-A' networks show more complementarity than similarity. Intuitively, words in 'Has-A' complement one another. For instance, 'a *house* has a *roof*' describes a complementary relation and these two objects are not similar to one another.
- Cluster 4 (dark blue/yellow): Most of the 'Related-To' and 'Union' networks show more similarity, except for Chinese. As defined in the 'Related-To' relation, words are connected if there is any sort of positive relationship between them, therefore triangles are easily formed. One exception to that rule is that the Chinese 'Related-To' network (dark blue cross) shows the strongest complementarity of all networks and lower-than-expected similarity. We find that a possible explanation is that the Chinese language has many measure words that are connected to a wide range of nouns. Measure words, also known as numeral classifiers, are used in combination with numerals to describe the quantity of things[70,71]. For example, the English 'one apple' translates to 一个苹果 in Chinese, where the measure word 个 (gè) must be added as a unit of measurement between the number 'one', 一 (yī), and the noun 'apple', 苹果 (píng guǒ). The Chinese measure word 个 can be loosely translated to English as 'unit(s) of', as in 'one unit of apple'. This grammatical construct is comparable to the phrase 'one box of apples' in English, where 'box' serves as a measure word, but, contrary to Chinese, measure words are rare in English. In the Chinese 'Related-To' network, there are many measure words that are connected to multiple nouns, and these nouns may have no connection with each other at all. Most of the measure words are not connected with each other either. Hence, the pairings of measure words and nouns lead to quadrangles, a likely explanation of why the Chinese 'Related-To' network shows the highest complementarity.

Discussion

In summary, we have conducted an exploratory analysis of the topological properties of semantic networks with 7 distinct semantic relations arising from 11 different languages. We identified both universal and unique characteristics of these networks.

We find that semantic networks are sparse and that many are characterized by a power-law degree distribution. We also find that many semantic networks are scale-free. We observe two different patterns of degree-degree mixing in these networks, some networks are assortative, while some are disassortative. In addition, we find that most networks are more clustered than expected.

On the other hand, some semantic networks – 'Related-To' in French, Spanish, Portuguese, and Finnish – have unique features that can be explained by rules of grammatical inflection. Because of the grammar in these languages, words have many conjugations or declensions. We have related anomalous peaks in the degree distributions to the language inflections. Notably, we found word inflection not only in inflecting languages but also in Finnish, which is an agglutinating language.

We have also quantified the relative roles of complementarity and similarity principles in semantic networks. The proportions of similarity and complementarity in networks differ depending on the semantic relation type. For example, the 'Synonym' networks exhibit stronger similarity, while the links in the 'Antonym' network are primarily driven by complementarity. In addition, the Chinese 'Related-To' network has the highest structural complementarity coefficient of all networks, which we attribute to a unique grammatical phenomenon in Chinese: measure words.

Through the analysis of the topological properties of semantic networks, we found that complementarity may play an important role in their formation. Since most of the state-of-the-art network inference methods are either built on or inspired by the similarity principle, we call for a careful re-evaluation of these methods when it comes to inference tasks on semantic networks. One basic example is the prediction of missing links. In a seminal work, Kovács et al.[64] demonstrated that protein interactions should be predicted with complementarity-tailored methods. We expect that similar methods might be in place for semantic networks. Instead of using the triangle closure principle, one might benefit from the methods based on quadrangle closure, Figure 7.

It is not as easy to illustrate quadrangle closure in network embedding methods or NLP methods in general. A plethora of methods use multiple modules and parameters in learning tasks and can, in principle, be better optimized for the complementarity structure of semantic networks. In our recent work, we propose a complementarity learning method and apply it to several networks, including the 'Antonym' semantic network[66].

Recent groundbreaking advances in large language models are attributed to the multi-head attention mechanism of the Transformer, which uses ideas consistent with the complementarity principle[72]. We advocate that a better understanding of statistical mechanisms underlying semantic networks can help us improve NLP methods even further.

Methods

Clustering coefficient.
We investigate clustering in semantic networks by measuring the clustering coefficient $c_G(i)$ of a node i, which equals the ratio of the number y of connected neighbors to the maximum possible number of connected neighbors,

$$c_G(i) = \frac{2y}{d_i(d_i - 1)},$$
(1)

as defined by Watts and Strogatz[33,42]. The graph clustering coefficient c_G is the average over all node clustering coefficients,

$$c_G = \frac{1}{N} \sum_{i=1}^{N} c_G(i).$$
(2)

We also calculate the average clustering coefficient $c_G(i)$ of nodes with degree $d_i = k$. In addition, we calculate $c_G(i)$ also after random rewiring for comparison.
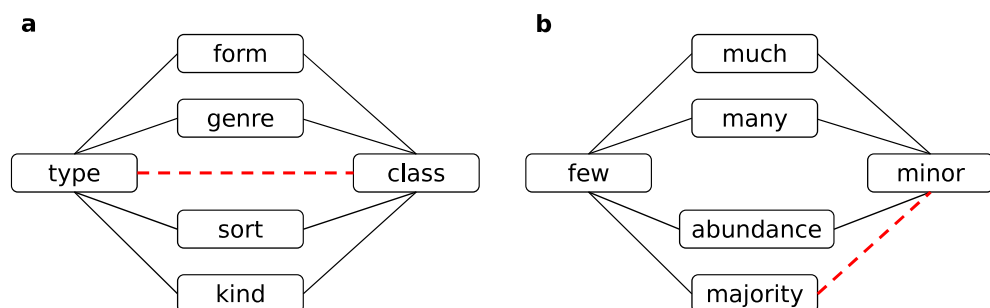
Figure 7. Examples of similarity and complementarity in semantic networks. **a** Similarity: triangle closure in the 'Synonym' network. **b** Complementarity: quadrangle closure in the 'Antonym' network.

Data Availibility

References

1. Cambria, E. & White, B. Jumping NLP curves: A review of natural language processing research. *IEEE Comput. Intell. Mag.* **9**, 48–57 (2014).
2. Quillian, M. R. Word concepts: A theory and simulation of some basic semantic capabilities. *Behav. Sci.* **12**, 410–430 (1967).
3. Quillian, M. R. The teachable language comprehender: A simulation program and theory of language. *Comm. ACM* **12**, 459–476 (1969).
4. Sowa, J. F. Semantic networks. *Encycl. Cognit. Sci.* **2**, 1429 (2012).
5. de Barros Pereira, H. B. *et al.* Systematic review of the "semantic network" definitions. *Expert Syst. Appl.* 118455 (2022).
6. Sowa, J. F. *Principles of semantic networks: Explorations in the representation of knowledge* (Morgan Kaufmann, 2014).
7. Peters, S. & Shrobe, H. Using semantic networks for knowledge representation in an intelligent environment. In *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003 (PerCom 2003)*, 323–329, https://doi.org/10.1109/PERCOM.2003.1192756 (2003).
8. Salem, A.-B. M. & Alfonse, M. Ontology versus semantic networks for medical knowledge representation. *Recent Adv. Comput. Eng.* 769–774 (2008).
9. Singhal, A. Introducing the knowledge graph: things, not strings. *Google Blog*https://blog.google/products/search/introducing-knowledge-graph-things-not *(2012).*
10. Popping, R. Knowledge graphs and network text analysis. *Soc. Sci. Inf.* **42**, 91–106 (2003).
11. Fensel, D. *et al.* Introduction: what is a knowledge graph? In *Knowledge Graphs*, 1–10 (Springer, 2020).
12. Kepuska, V. & Bohouta, G. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, 99–103 (IEEE, 2018).
13. High, R. The era of cognitive systems: An inside look at IBM Watson and how it works. *IBM Corp. Redb.* **1**, 16 (2012).
14. Piskorski, J. & Yangarber, R. Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, 23–49 (Springer, 2013).
15. Shi, F., Chen, L., Han, J. & Childs, P. A data-driven text mining and semantic network analysis for design information retrieval. *J. Mech. Des.* **139** (2017).
16. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
17. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Art. Intell. Res.* **11**, 95–130 (1999).
18. Harris, A. Human languages vs. programming languages. *Medium* https://medium.com/@anaharris/human-languages-vs-programming-languages-c89410f13252 *(2018).*
19. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Ann. Rev. Sociol.* 415–444 (2001).
20. Kossinets, G. & Watts, D. J. Origins of homophily in an evolving social network. *Am. J. Sociol.* **115**, 405–450 (2009).
21. Schaefer, D. R., Light, J. M., Fabes, R. A., Hanish, L. D. & Martin, C. L. Fundamental principles of network formation among preschool children. *Soc. Netw.* **32**, 61–71 (2010).
22. Snijders, T. A. Statistical models for social networks. *Ann. Rev. Sociol.* **37**, 131–153 (2011).
23. Hasan, M. A. & Zaki, M. J. A survey of link prediction in social networks. In *Social network data analytics*, 243–275 (Springer, 2011).
24. Zarandi, F. D. & Rafsanjani, M. K. Community detection in complex networks using structural similarity. *Phys. A: Stat. Mech.* **503**, 882–891 (2018).
25. Evans, N. & Levinson, S. C. The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–448 (2009).
26. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools. Appl.* 1–32 (2022).
27. Alfawareh, H. M. & Jusoh, S. Resolving ambiguous entity through context knowledge and fuzzy approach. *Int. J. Comp. Sci. Eng.* **3**, 410–422 (2011).
28. Jusoh, S. A study on NLP applications and ambiguity problems. *J. Theor. Appl. Inf. Technol.* **96** (2018).
29. Talaga, S. & Nowak, A. Structural measures of similarity and complementarity in complex networks. *Sci. Rep.* **12**, 16580 (2022).
30. Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation (1998).
31. Ferrer-i-Cancho, R. & Solé, R. V. The small world of human language. *Proc. R. Soc. B: Biol. Sci.* **268**, 2261–2265 (2001).
32. BNC Consortium. British National Corpus, XML edition. http://hdl.handle.net/20.500.12024/2554 (2007). Oxford Text Archive.
33. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
34. Motter, A. E., De Moura, A. P., Lai, Y.-C. & Dasgupta, P. Topology of the conceptual network of language. *Phys. Rev. E* **65**, 065102 (2002).
35. Ward, G. *Moby thesaurus list* (Quality Classics, 2015).
36. Sigman, M. & Cecchi, G. A. Global organization of the WordNet lexicon. *Proc. Nat. Acad. Sci.* **99**, 1742–1747 (2002).
37. Miller, G. A. WordNet: A lexical database for english. *Comm. ACM* **38**, 39–41 (1995).
38. Steyvers, M. & Tenenbaum, J. B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cogn. Sci.* **29**, 41–78 (2005).
39. Roget, P. M. *Roget's Thesaurus of English Words and Phrases* (TY Crowell Company, 1911).
40. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
41. Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
42. Van Mieghem, P. *Performance Analysis of Complex Networks and Systems* (Cambridge University Press, 2014).
43. Ferrer-i Cancho, R., Solé, R. V. & Köhler, R. Patterns in syntactic dependency networks. *Phys. Rev. E* **69**, 051915 (2004).
44. Noldus, R. & Van Mieghem, P. Assortativity in complex networks. *J. Complex Netw.* **3**, 507–542 (2015).
45. Speer, R., Chin, J. & Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4444–4451 (AAAI Press, 2017).
46. Voitalov, I., van der Hoorn, P., van der Hofstad, R. & Krioukov, D. Scale-free networks well done. *Phys. Rev. Res.* **1**, 033034 (2019).
47. Voitalov, I. Tail index estimation for degree sequences of complex networks. https://github.com/ivanvoitalov/tailestimation (2019).
48. Borge-Holthoefer, J. & Arenas, A. Semantic networks: Structure and dynamics. *Entropy* **12**, 1264–1302 (2010).

49. Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E. & Perc, M. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.* **2**, 943 (2012).
50. Adamic, L. A. & Huberman, B. A. Power-law distribution of the world wide web. *Science* **287**, 2115–2115 (2000).
51. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
52. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. *ACM SIGCOMM Comp. Commun. Rev.* **29**, 251–262 (1999).
53. Newman, M. E. Clustering and preferential attachment in growing networks. *Phys. Rev. E* **64**, 025102 (2001).
54. Lyons, J. Language classification. *Encyclopædia Britannica* https://www.britannica.com/science/linguistics/Other-relationships *(2022)*.
55. Eberhard, D. M., Simons, G. F. & Fennig, C. D. Ethnologue: Languages of the world. *SIL International* https://www.ethnologue.com *(2022)*.
56. Haspelmath, M. The morph as a minimal linguistic form. *Morphology* **30**, 117–134 (2020).
57. Speer, R. Relations in ConceptNet5. *ConceptNet 5 Wiki* https://github.com/commonsense/conceptnet5/wiki/Relations *(2019)*.
58. Comrie, B. *Aspect: An introduction to the study of verbal aspect and related problems* Vol. 2 (Cambridge University Press, 1976).
59. Kendris, C. & Kendris, T. *501 Spanish verbs* (Barrons Educational Series, 2020).
60. Vare, F. J. Your all-in-one guide to the 18 Spanish tenses and moods. *Enux Education, FluentU* https://www.fluentu.com/blog/spanish/spanish-tenses/ *(2022)*.
61. Nitti, J. J. & Ferreira, M. J. *501 Portuguese verbs* (Simon and Schuster, 2015).
62. Lawless, L. K. *The Everything French Verb Book: A Handy Reference for Mastering Verb Conjugation* (Simon and Schuster, 2005).
63. Karlsson, F. *Finnish: A comprehensive grammar* (Routledge, 2017).
64. Kovács, I. A. *et al.* Network-based prediction of protein interactions. *Nat. Commun.* **10**, 1–8 (2019).
65. Mattsson, C. E. *et al.* Functional structure in production networks. *Front. Big Data* **4**, 666712 (2021).
66. Budel, G. & Kitsak, M. Complementarity in complex networks. *arXiv preprint* arXiv:2003.06665 *(2020)*.
67. Rivera, M. T., Soderstrom, S. B. & Uzzi, B. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Ann. Rev. Sociol.* **36**, 91–115 (2010).
68. Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K. & Kivelä, M. Cumulative effects of triadic closure and homophily in social networks. *Sci. Adv.* **6**, eeax7310 (2020).
69. Jia, M., Gabrys, B. & Musial, K. Measuring quadrangle formation in complex networks. *IEEE Trans. Netw. Sci. Eng.* **9**, 538–551 (2021).
70. Tai, J. H. *Chinese classifier systems and human categorization. In honor of William S.-Y. Wang Interdisciplinary studies on language and language change* (Pyramid Press Taipei, 1994).
71. Cheng, L.L.-S. & Sybesma, R. Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua J. Chin. Stud.* **28**, 385–412 (1998).
72. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).

Acknowledgements

Author contributions

M.K. conceived the experiments, Y.J. conducted the experiments with support from G.B. and M.K, after which G.B., Y.J., P.V.M. and M.K. analyzed the results. All authors wrote, edited, and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-37294-8.

Correspondence and requests for materials should be addressed to M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.