Topological Properties of Semantic Networks

Ying Jin

Delft University of Technology

Topological Properties of Semantic Networks

by

Ying Jin

to obtain the degree of Master of Science
in Electrical Engineering
Track Wireless Communication and Sensing

at the Delft University of Technology,
to be defended publicly on Thursday September 29, 2022 at 11:00 AM.

Student number:	5184657	
Project duration:	November, 2021 – September, 2022	
Thesis committee:	Prof.dr.ir. Rob Kooij,	TU Delft, chair
	Dr. J.L.A. Dubbeldam,	TU Delft
	Dr. Maksim Kitsak,	TU Delft
Supervisors:	Prof.dr.ir P.F.A. Van Mieghem,	TU Delft
	Gabriel Budel,	TU Delft, daily supervisor

Cover image: https://www.istockphoto.com/

Preface

This thesis, ‘Topological Properties of Semantic Networks’, finalizes my Master of Science degree in Electrical Engineering at the Delft University of Technology. It has been an unforgettable experience for me to work on this project at the Network Architecture and Services (NAS) group.

During the past 9 months, I have grown huge interest in semantic networks and learnt much about structural properties of networks. First of all, my sincere gratitude to Professor Piet Van Mieghem for providing me with the opportunity to work on a fascinating topic. I would like to thank Professor Rob Kooij for his inspirational and helpful feedback. I would also like to express my great appreciation to Professor Maksim Kitsak for guiding me through the research with constant motivation and supportive feedback. I would not have achieved the success of this thesis without his expertise and selflessness. Also many thanks to my daily supervisor, Gabriel Budel, for giving me excellent ideas, advice and tips throughout the thesis. It has been a great fortune for me to work with him.

Additionally, I would like to thank Professor Johan Dubbeldam to be on my thesis committee. Studying my master degree during the Covid period has been a challenge. I would like to thank the supportive and lovely NAS group for having me as a member. I will never forget this pleasant journey that has been filled with inspiration, joy and of course, knowledge.

Lastly, a huge amount of thanks to my family and friends for their unconditional love, company and support. I cannot say enough about how lucky I am to have them, for they are my courage and strength to keep on exploring the universe.

Ying Jin
Rotterdam, September 2022

Abstract

The main goal of this thesis is to understand the topological properties of semantic networks, to find language-specific patterns, and to investigate their connection principles. Interpreting unstructured texts in natural language is a crucial task for computers. Natural Language Processing (NLP) applications rely on semantic networks for structured knowledge representation. Although NLP technologies have been applied to various domains with some degree of success, they still face many challenges due to the ambiguity of human language. To inform better algorithms, we need to pay attention to fundamental structures of semantic networks in different languages. However, these remain to be investigated properly. In this thesis we extract semantic networks with 7 distinct relations for 11 languages from ConceptNet. We systematically analyze the degree distribution, degree correlation and clustering of these networks. We also measure their structural similarity and complementarity coefficients. Our findings show that semantic networks have universalities in basic structures: they have high sparsity, high clustering, and power-law degree distributions. Our findings also show that the majority of the considered networks are scale-free. In addition, our results show that networks in different languages exhibit different properties, which are determined by grammatical rules. For example, the networks of highly inflected languages show peaks in the degree distributions that deviate from a power-law. Furthermore, we find that depending on the type of semantic relation and the language, the connection principles of networks are different. Some networks are more similarity-based, while others are more complementarity-based. We conclude the thesis by demonstrating how the knowledge of similarity and complementarity can better inform NLP in link prediction tasks.

Contents

1
Introduction

Due to the explosion in availability of digital content over time, the demand for computers to efficiently handle textual data has never been greater. The large amounts of data and computing power have enabled a significant amount of research on Natural Language Processing (NLP). The goal of NLP is to allow computer programs to interpret and process human language texts. A text is represented in a computer as a string, but human language is much more than just a string. We can relate various concepts to a text based on our 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 play an important role in representing human knowledge.

A semantic network is a graph representation of structured knowledge. It is composed of nodes, which represent concepts (e.g., words or phrases), and links, which represent semantic relations between concepts [2, 3]. 'Semantic' means 'relating to meaning in language or logic'. Fig. 1.1 presents a toy example of a semantic network. In the 1960s, semantic networks were first suggested by Quillian [4, 5] as a means of representing human knowledge in a computer.

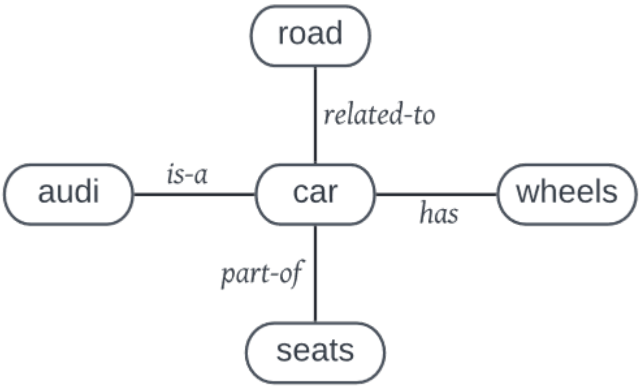


Figure 1.1: Toy example of a semantic network with five concepts and four semantic relations.

The past two decades have witnessed a rise in the significance of semantic networks in NLP [6–8]. Today, activities in our daily lives are already inseparable from semantic network applications. Semantic networks are employed in NLP technologies to represent and extract knowledge. 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]. The use of Knowledge Graph allows Google's computers to store and analyze complicated information, improving search results and user experience. Likewise, 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 the received information, comprehend the conversation, and achieve the goal of communicating with users.

1.1. Motivation and Objectives

Language is a complex system with diverse grammatical rules. To grasp the meaning of a sentence, humans count on their natural understanding of language and concepts in contexts. However, it is difficult for computers to utilize similar strategies since computers do not have the capability of understanding. 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 rules are ambiguous [16]. After all, in computer languages, there are no synonyms, namesakes or tones that can lead to misinterpretation [17]. Thus, computers rely on external tools to enable the processing of the structure and meaning of texts.

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

- *Inform better NLP methods.*
  Although there have been numerous real-world NLP applications across various domains, existing NLP technologies still face limitations [18]. For example, processing texts in a language where multiple words have similar meanings make a difficult task for computers [19]. Other issues occur due to lexical ambiguity when a word or phrase conveys more than one meaning [20, 21]. Furthermore, it is challenging to analyze a text that includes spelling inconsistencies, dialects or culture-specific phrases [22, 23]. Existing algorithms are usually domain-specific, and achieving more accurate and broader applications remains a problem. To design finer language models that can handle NLP challenges such as language ambiguity, we first need to understand what semantic networks look like. Therefore, it is necessary to study the topological properties of semantic networks.

- *Understand fundamental formation principles of semantic networks.*
  In many social networks such as friendship networks, connections between nodes are driven by similarity [24–27]. The more similar (in terms of number of common neighbors) two nodes are, the more likely it is that 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 [28] and community detection [29]. But these tools may not work well with 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 would like to learn what the basic principles are that drive the formation of semantic networks.

- *Document language-specific features.*
  Languages vary greatly between cultures and across time [30]. Two languages that originate from two different language families can differ in many types of features since they are structured based on different rules. It is natural to conjecture that there exist diverse structures in semantic networks for different languages.

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

To sum up, the main objective of this thesis 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. Secondly, 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.

1.2. Contributions

The main contributions of this thesis include:

1. We extract semantic networks based on semantic relations (link types). We study topological properties of seven English semantic networks. Each network is defined by a specific link type. We show that all networks possess high sparsity and a power-law degree distribution. In addition, we find that most networks have a high average clustering coefficient, while others show the opposite.

2. We extend the study of the topological properties of semantic networks to ten other languages. We perform analyses based on two types of language classifications. 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 degree distribution and results in deviations from a power-law distribution. We find this feature not only in inflecting languages, but also in one language that is classified as agglutinating.

3. We study the organizing principles of 50 semantic networks. We apply the algorithm from [31] to quantify the structural similarity and complementarity of semantic networks, and show to what extent these networks are similarity- or complementarity-based. Furthermore, we present that the connection principles in semantic networks are related to the type of semantic relation.

1.3. Thesis Outline

This thesis is organized in the following manner: In Chapter 2 we first introduce fundamental concepts from graph theory and complex networks on which this thesis relies. Additionally, some network models and randomization methods are presented. Then, we provide a brief overview of recent work on semantic networks. Finally, we conclude the chapter by clarifying the research gap and the scope of our research. In Chapter 3 we present the general topological properties of seven English semantic networks. We also introduce our dataset, types of semantic relation and network extraction procedure. In Chapter 4 we study semantic networks from different languages. By inspecting the degree distributions, we find patterns related to grammatical inflections in several languages. We also compare topological properties between multiple language families. Chapter 5 deals with the fundamental connection principles in semantic networks in eleven languages. We measure and compare the structural similarity and complementarity of different networks. We present and discuss the patterns that we find. Finally, we draw conclusions from the results and findings obtained in the thesis and give recommendations for future research in Chapter 6.

2 Background

A semantic network is a graph that represents human knowledge. Before diving into the study of topological properties of semantic networks, we introduce the terminologies in graph theory and properties of complex networks that will be utilized in the subsequent chapters. Besides, we explain the random network models that are used and two methods of network randomization. After clarifying network-related concepts, we briefly discuss recent work on semantic networks in the last section. At last, we refine the objective and scope of this thesis.

2.1. Graph Theory

A network is a **graph** $G(N, L)$ that consists of N number of nodes (vertices) and L number of links (edges). Nodes are connected via links. A **self-loop** is a link that has the same node as the endpoints. A **subgraph** of a graph G is a graph whose nodes and links all belong to G [32].

A network is called **undirected** if the links do not specify the source and destination nodes, otherwise the network is **directed**. Networks with self-links that have the same source and destination nodes are not taken into account in this thesis. Sometimes, links are associated with weights. We call this type of network a **weighted** network. In this thesis, we only consider *undirected* and *unweighted* networks. We provide explanations for the choice in the end of this chapter.

In an undirected network, a **path** is a sequence of links that joins a set of nodes in which all nodes are distinct [32]. The **length** of a path equals the number of links in the path. For example, a path of length 2 that originates from node i and travels via j to k can be represented as (i, j, k).

An **adjacency matrix** A is a matrix representation of a network that provides the complete link information. It is an $N \times N$ matrix where every element a_{ij} is equal to 0 or 1. If a link exists between node i and j, then $a_{ij} = 1$; otherwise, $a_{ij} = 0$.

2.2. Topological Properties in Complex Networks

The primary metrics that we use in this thesis are listed in Table 2.1 together with their mathematical notations. We consider the following metrics: maximum degree d_{max}, average degree $E[D]$, degree distribution $\Pr[D = k]$, degree correlation coefficient

ρ_D, Average Nearest Neighbor Degree (ANND), clustering coefficient c_G and graph transitivity $\check{c}_G$.

Mathematical notation	Metrics
N	number of nodes
L	number of links
k	degree
d_{max}	maximum degree
$E[D]$	average degree
$\Pr[D = k]$	degree distribution
ρ_D	degree correlation coefficient
ANND	average nearest neighbor degree
c_G	clustering coefficient
$\check{c}_G$	graph transitivity
γ	power-law exponent for the degree distribution

Table 2.1: Primary metrics used in the thesis and their mathematical notations.

2.2.1. Degree

We begin with the simplest, yet key, property of a network: the node degree. The **degree** d_i of a node i in a graph $G(N, L)$ equals the number of its **neighbors**, *i.e.*, the number of links that connect to node i [33]. The degree of node i satisfies $0 \leq d_i \leq N - 1$. The maximum degree $d_{max} = N - 1$ is achieved in a connected graph for nodes with the most neighbors. The sum of nodal degrees in a network follows the rule given by Eq. 2.1, which is twice the number of links

$$\sum_{i=1}^{N} d_i = 2L. \tag{2.1}$$

The **average degree** $E[D]$ of a network is an important global measure, it is defined as

$$E[D] = \frac{1}{N}\sum_{i=1}^{N} d_i = \frac{2L}{N}. \tag{2.2}$$

In network theory, degree distribution plays a crucial role. Even though the degree of individual nodes is a local metric, the distribution of degrees provides a global view of the structure of a network. The density of the **degree distribution** $\Pr[D = k]$ is the fraction of nodes in a network that have degree k. Mathematically,

$$\Pr[D = k] = \frac{N_k}{N}, \tag{2.3}$$

where N_k is the number of $d = k$ nodes, and N is the total number of nodes. The degree distribution shows the probability that a network node chosen at random has degree k [34]. The probability $\Pr[D = k]$ is normalized as

$$\sum_{k=1}^{N-1} \Pr[D = k] = 1. \tag{2.4}$$

It is useful to plot the degree distribution as a function of degree k, using a histogram or scatter plot. The shape of the curve illustrates important structure properties of various kinds of networks. The degree distribution of many real complex networks is a **power-law** distribution [35]

$$\Pr[D = k] \sim k^{-\gamma}, \tag{2.5}$$

where γ is the power-law exponent. By taking the logarithm of Eq. 3.1, we have

$$\log(\Pr[D = k]) \sim -\gamma \log(k). \tag{2.6}$$

If we plot the logarithm of degree distribution $\log(\Pr[D = k])$ with respect to the logarithm of degree $\log(k)$, we should see a linear dependency between $\log(\Pr[D = k])$ and $\log(k)$. And the slope of the line is the power-law exponent γ.

For **scale-free** networks, the power-law exponent γ typically lies between 2 and 3. The power-law exponent $\gamma \in (2, 3)$ means the average degree is finite, but the variance is infinite.

2.2.2. Degree Assortativity

Degree assortativity, also known as assortative mixing, describes the tendency of nodes to connect to other nodes with either similar or opposite degree [36]. A network is said to be assortative if high-degree nodes connect to high-degree nodes. Conversely, a network is disassortative if high-degree nodes connect to low-degree nodes. There are two common measures for capturing degree assortativity.

The first metric is the **degree correlation coefficient** ρ_D, defined as the linear correlation coefficient of the degrees at either ends of a link $l = i \sim j$ [36]. Van Mieghem [33] expresses the degree correlation coefficient ρ_D in terms of graph metrics as

$$\rho_D = 1 - \frac{\sum_{i\sim j}\left(d_i - d_j\right)^2}{\sum_{i=1}^{N} d_i{}^3 - \frac{1}{2L}\left(\sum_{i=1}^{N} d_i{}^2\right)^2}, \tag{2.7}$$

where d_i and d_j are the degrees of node i and j which are connected by a link. If $\rho_D > 0$, the network is assortative. And $\rho_D < 0$ denotes a disassortative network.

The other metric is **Average Nearest Neighbor Degree** (ANND) as a function of the degree k. It is determined by the number of nodes and the degrees of its direct neighbors in a network. The definition of the ANND as a function of the degree was introduced by Boguñá and Pastor-Satorras [37]. The average nearest neighbor degree of node i, ANND(i), is the expected degree of all its neighboring nodes. We rewrite the definition of ANND(i) using the adjacency matrix as follows

$$\text{ANND}(i) = E[D_j | a_{ij} = 1], \tag{2.8}$$

where D_j is the degree of a neighboring node j. Therefore, the average nearest neighbor degree *ANND* of the whole network is the average over all nodes

$$\text{ANND} = \frac{1}{N}\sum_{i=1}^{N} \text{ANND}(i). \tag{2.9}$$

Similarly, we can calculate ANND for nodes with degree k. Then we will have a function of ANND with respect to k, which we denote as ANND(k). A function that explicitly depends on degree k indicates the existence of degree correlations in a network. According to Newman [36], when ANND(k) is an increasing function of k, the network is assortative. Contrariwise, a network is disassortative if ANND(k) decreases along k.

2.2.3. Clustering

The **clustering coefficient** quantifies the graph connectivity structure. Here we introduce three basic measures of clustering.

The first one is the clustering coefficient $c_G(i)$, which quantifies the local density around a node i. It is defined as the ratio of the number of connected neighbor pairs of node i over the number of its all possible neighbor pairs [38, 39],

$$c_G(i) = \frac{\sum_{j,k} a_{ij} a_{ik} a_{jk}}{d_i(d_i - 1)}. \tag{2.10}$$

The local clustering coefficient $c_G(i)$ ranges from 0 to 1. When there are zero connection between the neighbors of node i or the degree of the node is less than or equal 1 ($d_i \leq 1$), $c_G(i) = 0$. The maximum $c_G(i) = 1$ is reached only when the neighbors of node v are all connected.

The second measure is the average clustering coefficient c_G of an entire network. It measures how closely nodes in a network cluster together. The clustering coefficient c_G of a network with N nodes is defined as the average over all nodes

$$c_G = \frac{1}{N} \sum_{i=1}^{N} c_G(i). \tag{2.11}$$

Similar to the local clustering coefficient $c_G(i)$, the average clustering coefficient satisfies $0 \leq c_G \leq 1$. The minimum $c_G = 0$ is attained in graphs where there are no triangles at all, while the largest $c_G = 1$ only happens when a network is fully connected.

The **graph transitivity** $\check{c}_G$, often referred to as the global clustering coefficient [38] or the ratio of transitive triples [40], calculates the ratio of the number of closed triangles in a network relative to the total possible triples. As defined in [40, 41],

$$\check{c}_G = \frac{6t}{\sum_{i=1}^{N} d_i(d_i - 1)}, \tag{2.12}$$

where t is the total number of triangles in a network. The value of the graph transitivity $\check{c}_G$ is constrained within 0 and 1. A fully connected graph has $\check{c}_G = 1$, while $\check{c}_G$ goes to zero for a random graph as its size increases.

2.2.4. Connectedness

A graph is **connected** if there exists a path between any pair of its nodes [42]. Connectedness is an important concept in graph theory. In a not fully connected graph, a **connected component** is a connected subgraph that does not belong to any other connected subgraph of a bigger size..

The **Largest Connected Component** (LCC) is the connected component with the largest number of nodes in a network. If a network is connected, then the network itself is the largest connected component.

In this thesis, we focus on the study of the largest connected component of semantic networks.

2.3. Random Networks

A random network is a network in which the connections are completely random apart from specific constraints. One famous random network model is the *Erdős–Rényi random graph* $G(N, L)$, where N nodes are randomly connected with L links [43]. In such a network, the average degree is fixed to $E[D] = 2L/N$. A more flexible network model is the configuration model. The *Configuration Model* (CM) allows users to generate networks with any desired degree sequence [44]. As a consequence, the degree of each node is fixed. However, networks created using the configuration model may contain self-loops and multi-links. A multi-link is composed of more than one link that have two identical endpoints.

In many statistical analyses, results obtained from randomized networks serve as benchmarks for comparison with real networks. Hence, we introduce two algorithms to randomize the semantic networks. The algorithms are *degree-preserving network rewiring* and *degree-preserving network reconstruction*.

The purpose of degree-preserving randomization is to keep the link density of each node but randomize the connections among all nodes. This way, we can compare the extracted real networks with the random networks that are expected by chance based on node degrees. **In this thesis, the *rewired* and *reconstructed* networks are obtained via these two randomization methods respectively.**

2.3.1. Degree-preserving Network Rewiring

Degree-preserving network rewiring randomly rewires the links between nodes, but keeps the degree of all nodes unchanged. To preserve the degrees of all nodes, we randomly select 1 link pair (4 nodes), and swap the endpoints of these 2 links. Fig. 2.1 illustrates the rewiring method. To make sure that all links are likely to be rewired at least once, we repeat the random selection of links for T times, where T is set to four times the number of links. The pseudocode is provided in Algorithm 1.

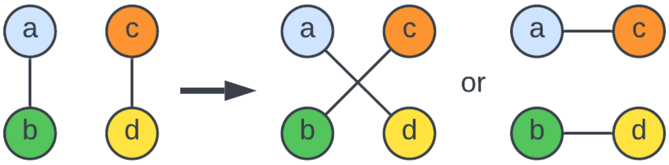


Figure 2.1: Illustration of degree-preserving rewiring. By randomly swapping the endpoints of two links (a, b) and (c, d), new links can be constructed without changing the node degrees.

Algorithm 1: Degree-preserving network rewiring

```
Data: a list of links
Result: a rewired network
$E$ ← a list of links;
$T$ ← $4L$ ;              /* all links are rewired at least once */
while $T \neq 0$ do
    $(a, b)$ and $(c, d)$ ← randomely pick 2 links from $E$;
    $n$ ← |set$(a, b, c, d)$| ;  /* number of unique nodes in 2 links */
    if $n < 4$ then
        continue
    else
        $a$ and $c$ ← randomly select one node from each link ;
        $(c, b)$ and $(a, d)$ ← swap the two slected nodes;
        if $(c, b) \in E$ or $(a, d) \in E$ then
            continue
        else
            $E$ ← update the list of links with the 2 rewired links $(c, b)$ and $(a, d)$;
            $T \leftarrow T - 1$
        end
    end
end
```

2.3.2. Degree-preserving Network Reconstruction

In degree-preserving network reconstruction, all links in a network are disconnected, then links are added randomly between nodes according to the original degree of each node, until all degrees are satisfied. We use a weighted probability random number generator to select nodes such that the probability of a node being chosen is proportional to its degree. The pseudocode is provided in Algorithm 2.

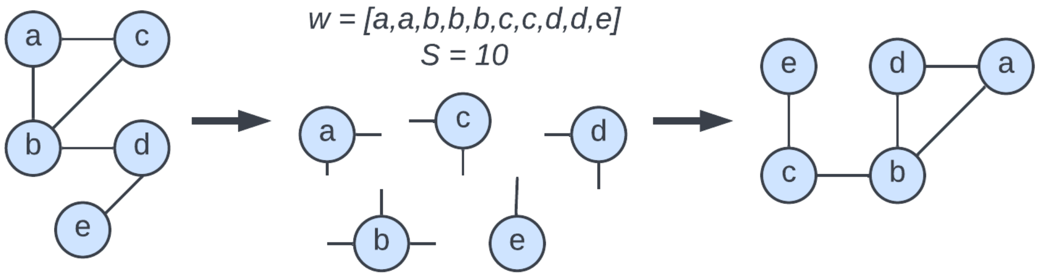


Figure 2.2: Illustration of degree-preserving reconstruction. After disconnecting all nodes, every node is left with an unmatched degree number. In this example, the array w has a length of 10. And the sum of all unmatched degree values $S = 10$.

First, we create two variables. The first variable is an array **w** with length $2L$, which is the degree sum of all nodes. The elements inside the array are node labels i repeated d_i times, where d_i is the degree of node i. The second one is a variable S that indicates the number of unmatched node degrees, the initial value of S equals the length of the array **w**. An example is given in Fig. 2.2. The resulting networks have degree distributions very close to the original ones. Specifically, in some instances, only a few nodes are left with degree value less than the original ones.

Algorithm 2: Degree-preserving network reconstruction

```
Data: a list of links
Result: a reconstructed network
$E$ ← an empty list storing links;
$w$ ← an array of node labels (weighted by node degrees, see Fig. 2.2);
$S \leftarrow 2L$;
while $S \neq 0$ do
    $a$ and $b$ ← randomely pick 2 nodes from $w$;
    if $a = b$ then
        continue
    else
        if link $(a, b) \in E$ then
            continue
        else
            $E$ ← add link $(a, b)$ to the list $E$;
            $w$ ← remove the chosen node labels $a$ and $b$ from $w$;
            $S \leftarrow S - 2$
        end
    end
end
```

2.4. Related Work

With the prior knowledge of complex networks, we now review some important work that has been carried out concerning the structure of semantic networks. Due to the vast interest in semantic networks, the related studies were carried out in different fields for diverse purposes. Based on our scope, we concentrate on two main aspects: (1) Topological properties that were analyzed in the literature, including the dataset used. (2) Common or different patterns in different languages which were found and discussed.

The majority of semantic networks literature targeted at three link types:

- *Co-occurrence*
  In a co-occurrence network, a pair of words that co-occur in a sentence or text form a link. This is commonly used in text analysis.

- *Association*
  A word can be associated with multiple other words. In cognitive-linguistic experiments, participants are given a word and asked to give the first word that they think of. There are several Free Association data sets, one example is the University of South Florida Free Association Norms [45].

- *Semantic relation*
  Semantic relations are defined by professionals like lexicographers. Typical semantic relations are synonym, antonym, hypernym and homonymy. In networks constructed from dictionary or thesaurus, semantic relations are the links that connect one word to another.

In 2001, Ferrer-i-Cancho and Sole [46] studied undirected co-occurrence graphs constructed from the British National Corpus dataset [47]. They measured the average distance between two words and observed the small-world property, which was found in many natural networks [38]. Motter *et al.* [48] analyzed an undirected conceptual network constructed from an English Thesaurus dictionary [49]. They focused on three properties: sparsity (small average degree), average shortest path length and clustering. That same year, Sigman and Cecchi [50] studied undirected lexical networks extracted from WordNet [51, 52], where nodes are various noun meanings. 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. Both [48] and [50] highlighted that semantic networks possess small-world structure with sparse connectivity, short average path lengths, and strong local clustering.

Later, Steyvers and Tenenbaum [53] performed statistical analysis of 3 kinds of semantic networks: word associations [45], WordNet and Roget's Thesaurus [54]. Apart from the above mentioned network properties, they also considered network connectedness and diameter. They pointed out that the small-world feature may origin from the scale-free organization of the network, which exists in a variety of real-world systems [35, 55].

As for patterns in different languages, Ferrer-i-Cancho *et al.* [56] built syntactic dependency networks from corpora (collections of sentences) for three European languages: Czech, German, and Romanian. They showed that networks in different languages have many non-trivial topological properties in common, such as small world structure, power-law degree distribution and disassortative mixing.

Existing studies have obtained general network structures such as small-world structure and power-law degree distribution. However, the text sources used for building these semantic networks are heterogeneous. Some used associative networks generated from experiments, and some chose a thesaurus manually created by people. Plus, most of the research performed coarse-grained statistical analyses. Specifically, various semantic relations were treated equally and nodes were only words (some only certain type of words, like nouns). Further, there are only very few studies on semantic networks of languages other than English.

Therefore, our analyses focus on semantic networks with distinct semantic relations (link types). We look at **undirected** and **unweighted** networks with specific link types, and compare the structural properties among networks with different link types. In addition, we apply similar analyses to such defined networks across different languages. Furthermore, we investigate how similarity and complementarity play roles in the connection principles of semantic networks.

3

General Topological Properties of Semantic Networks

To understand the structure of semantic networks, we start with their general topological properties. In this Chapter, we first introduce the dataset that we use throughout the thesis. Based on the primary objectives of the study, we provide the reasoning for our choice over other available datasets. Next, we define a number of semantic relations that are adopted to construct our networks. Later, we explain the procedure for building our semantic networks from the dataset, including the data cleaning process. At the moment, we center on the English semantic networks. Then, we compute various topological properties of these networks related to the connectedness, degree, assortativity and clustering. We characterize the structure of semantic networks with the obtained quantitative results. Finally, we give a summary of the overall network statistics.

3.1. Selection of the Dataset

There are a variety of datasets that can be used for semantic network analysis. Because this thesis focuses on semantic networks with different relations and languages, we draw our attention to the large datasets (*i.e.*, the total number of words is at least larger than 5k). Three frequently mentioned datasets in the network literature are listed as follow:

- *WordNet* [51] is a lexical database that was collected manually by lexicographers. It resembles a thesaurus, where nodes that have similar meanings are grouped together [57]. WordNet stores data in ASCII format across several files. The number of words ranges from 5k to 100k. It contains three types of relations: hyperonym, meronym and antonym. However, this dataset is only available in English.

- *DBPedia* [58] is an open knowledge graph that stores structured information extracted from Wikipedia. It was created by crowdsource groups from various projects. This large multilingual database consists of links connecting website pages. There are around 14 million links in DBPedia, but the link types (semantic relations) are limited.

- *ConceptNet* [59] is a multilingual database in the form of a semantic network where nodes are words and phrases from natural language. Links indicate in total 34 semantic relations. The knowledge is collected from a variety of resources, including crowdsourced resources, expert-created resources, and games with a purpose [59]. Most languages have more than 200k different nodes.

Based on the difficulty of network extraction, the number of specified semantic relations and the size of the dataset, we decided to use ConceptNet. The structure of ConceptNet is in the form of assertions, which are units of knowledge in ConceptNet. An assertion contains multiple entries, such as Uniform Resource Identifier (URI), language, relation, sources and weight. A URI specifies the relation between two natural-language concepts and what languages they are in. This is convenient for the network extraction. Besides, there are 34 defined relations that connect the nodes of ConceptNet [60]. ConceptNet covers hundreds of languages, among which 78 of them have at least 10k concepts (words or phrases). Moreover, ConceptNet includes parts of WordNet and DBPedia.

3.2. Semantic Relations

We consider in total 7 link types (relations), and 6 of them are selected from ConceptNet. They are 'Has-A', 'Part-Of', 'Is-A', 'Related-To', 'Antonym' and 'Synonym'. These are most meaningful and important relations with sufficiently large data. In addition, we define an additional link type 'Union', which is the integration of four networks, 'Has-A', 'Part-Of', 'Is-A' and 'Related-To'. The purpose of adding this link type is to treat all four relations equally and to see how the structure of the whole network is different from the individual ones. Table 3.1 shows the definition of selected six relations and related examples from ConceptNet.

Relation	**Description**	**Directed**	**Examples**	**Creation Method**
Has-A	B belongs to A, either as an inherent part or due to a social construct of possession. Has-A is often the reverse of Part-Of.	Yes	bird → wing	Manual + Automatic
Part-Of	A is a part of B. This is the part meronym relation in WordNet.	Yes	gearshift → car	Manual + Automatic
Is-A	A is a subtype or a specific instance of B; every A is a B. This can include specific instances; the distinction between subtypes and instances is often blurry in language. This is the hyponym relation in WordNet.	Yes	car → vehicle	Manual + Automatic
Related-To	The most general relation. There is some positive relationship between A and B, but ConceptNet can't determine what that relationship is based on the data.	No	learn ↔ erudition	Manual + Automatic
Antonym	A and B are opposites in some relevant way, such as being opposite ends of a scale, or fundamentally similar things with a key difference between them. Counterintuitively, two concepts must be quite similar before people consider them antonyms. This is the antonym relation in WordNet.	No	black ↔ white	Automatic
Synonym	A and B have very similar meanings. They may be translations of each other in different languages. This is the synonym relation in WordNet.	No	sunlight ↔ sunshine	Automatic

Table 3.1: Definition of the six relations and related information from ConceptNet [60].

Though some networks are directed, i.e., 'Has-A', 'Part-Of' and 'Is-A', we treat all networks as undirected. In most of the literature on network topologies, undirected networks were studied as key benchmarks mainly because the methods were developed for undirected networks only. To avoid sophisticated analysis of directed networks, and to meet the requirements of future work, we focus on undirected networks.

3.3. Extraction of Semantic Networks

Processing the dataset and extracting the desired semantic networks is the preparation of our network analysis. We first explain the extraction process of English semantic networks with different link types. In ConceptNet, a complete link information exists in the Uniform Resource Identifier (URI) provided by an assertion. Due to the clear structure, we are able to identify both the two natural-language concepts (nodes) and the relation between them. Besides, the languages of two concepts are also indicated. First, we extract the English subgraph with all link types from ConceptNet. Then, for each relation (link type), we extract links from the obtained English subgraph and store the networks in the edge list format. An edge list is a network representation in the form of edges, where each edge contains a start node and an end node. Each link is unweighted, undirected and connects two different nodes. For the convenience of statistical analysis, we give numerical labels to the nodes in a network and save them in an edge list. At the same time, we keep an index file for every semantic network, which indicates the nodes (*i.e.*, in words) and corresponding labels.

3.3.1. Multi-word Phrases

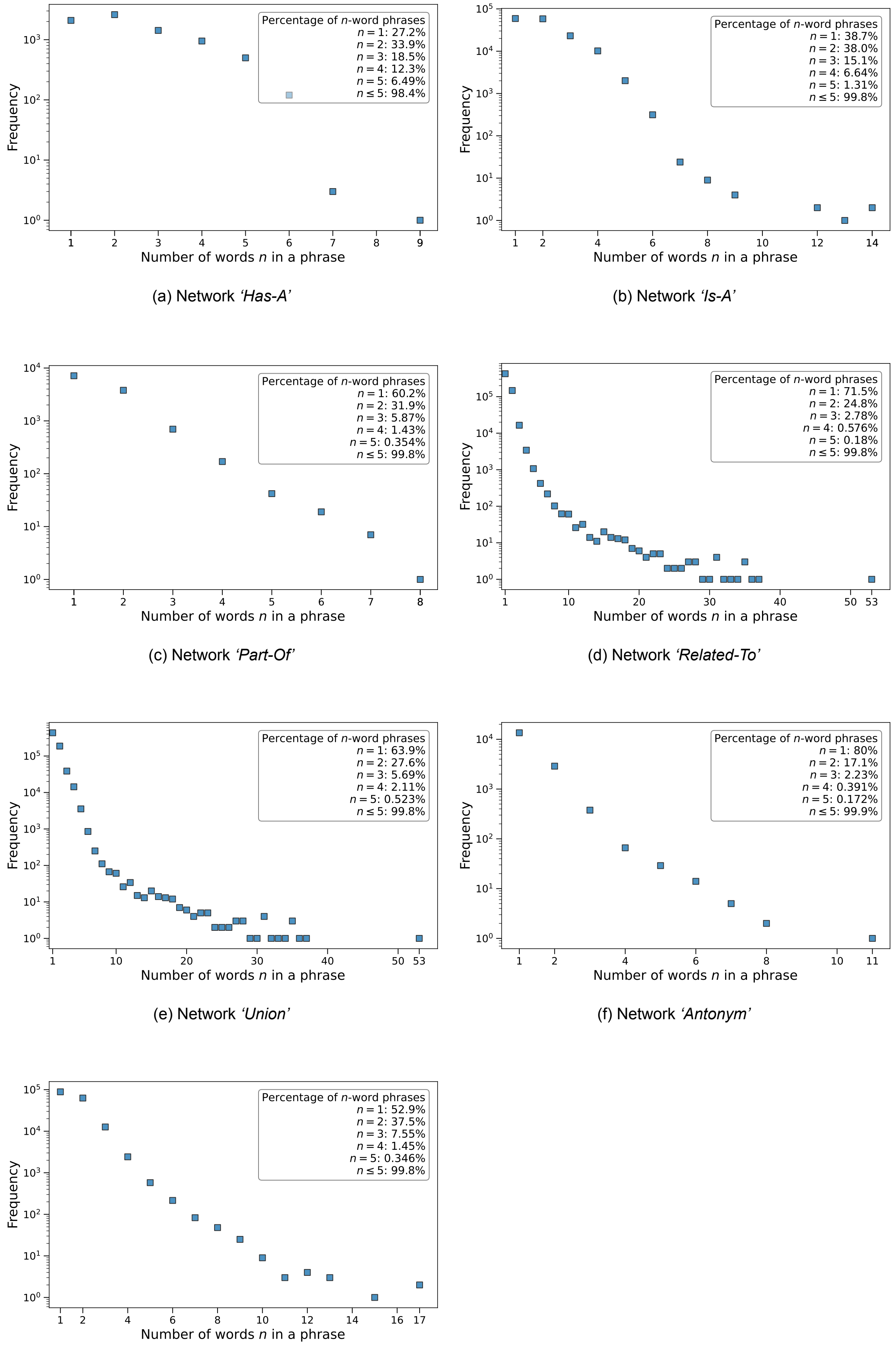
In our networks, a phrase is represented by multiple words that are connected by underscores. During the extraction, we discover that some phrases contain many words. To understand the composition of the networks better, we count the occurrences of phrases with different numbers n of words for each network. For example, the phrase ‘a_lot_of_places’ has $n = 4$ words. Note that when $n = 1$, the phrase is equivalent to a word, such as ‘cat’, ‘plant’ and ‘people’.

Fig. 3.1 reveals the frequencies of phrases with $n = 1, 2, ..., 5$ words in the seven networks. In the upper right corner of the plots, we present the percentage of n-word phrases in entire networks as well as the total percentages of phrases with $n \leq 5$. It is clear that phrases with not more than 5 words make up almost the entire (>98%) network. In addition, words ($n = 1$) are the major type of nodes most networks (except for the ‘Has-A’ network). Fig. 3.1a shows that ‘Has-A’ network has more 2-word phrases than words.

The maximum phrase length n_{max} of each network is listed in Table 3.2. Most networks have phrases with the maximum phrase length n_{max} less than 20. However, it is surprising to see that in some networks, *i.e.*, ‘Related-To’ and ‘Union’, the maximum phrase length n_{max} reaches 53 words. Such long phrases are usually not common and lack practical meaning, and they are likely caused by automatic extraction part of the dataset.

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
n_{max}	9	14	8	53	53	11	17

Table 3.2: The maximum number of words n_{max} in a phrase of the seven networks.



(a) Network *'Has-A'*

(b) Network *'Is-A'*

(c) Network *'Part-Of'*

(d) Network *'Related-To'*

(e) Network *'Union'*

(f) Network *'Antonym'*

(g) Network *'Synonym'*

Figure 3.1: Histogram of n-word phrases in the seven English semantic networks extracted from ConceptNet. In the upper right corner of the plots, we present the percentage of n-word phrases in entire networks as well as the total percentages of phrases with $n \leq 5$.

After investigating the distributions of n-word phrases of all networks, we arrive at a conclusion that the raw semantic networks need further processing. Specifically, filtering longer phrases is necessary. Since the networks are almost entirely comprised of phrases with less than 5 words, intuitively, we set the cut-off value n equals 5 for filtering all seven networks. The uniform cut-off value is for the consistency in nodes among all networks.

3.4. Overview of Semantic Networks

We calculate the overall descriptive statistics of the seven semantic networks, they are the number of nodes N, the number of links L, the maximal degree d_{max} and the average degree $E[D]$. Table 3.3 summarizes the results.

Based on the number of nodes, network 'Has-A' is the smallest and 'Union' is the largest. Given the network sizes, all of them have a very small average degree. For instance, in network 'Part-Of', on average a node only has connections to 2 (0.02%) of the 11,839 total number of nodes. *In other words, the number of links are of the same order as the number of nodes, which indicates that semantic networks are sparse.* With this in mind, we continue to study other topological properties in the following sections.

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	7,503	152,538	11,839	592,816	677,426	16,867	166,922
L	5,421	220,589	12,003	1,610,452	1,819,646	14,371	155,048
d_{max}	372	2913	116	4025	5263	38	103
$E[D]$	1.45	2.89	2.03	5.43	5.37	1.70	1.86

Table 3.3: Basic statistics of the seven English semantic networks extracted from ConceptNet.

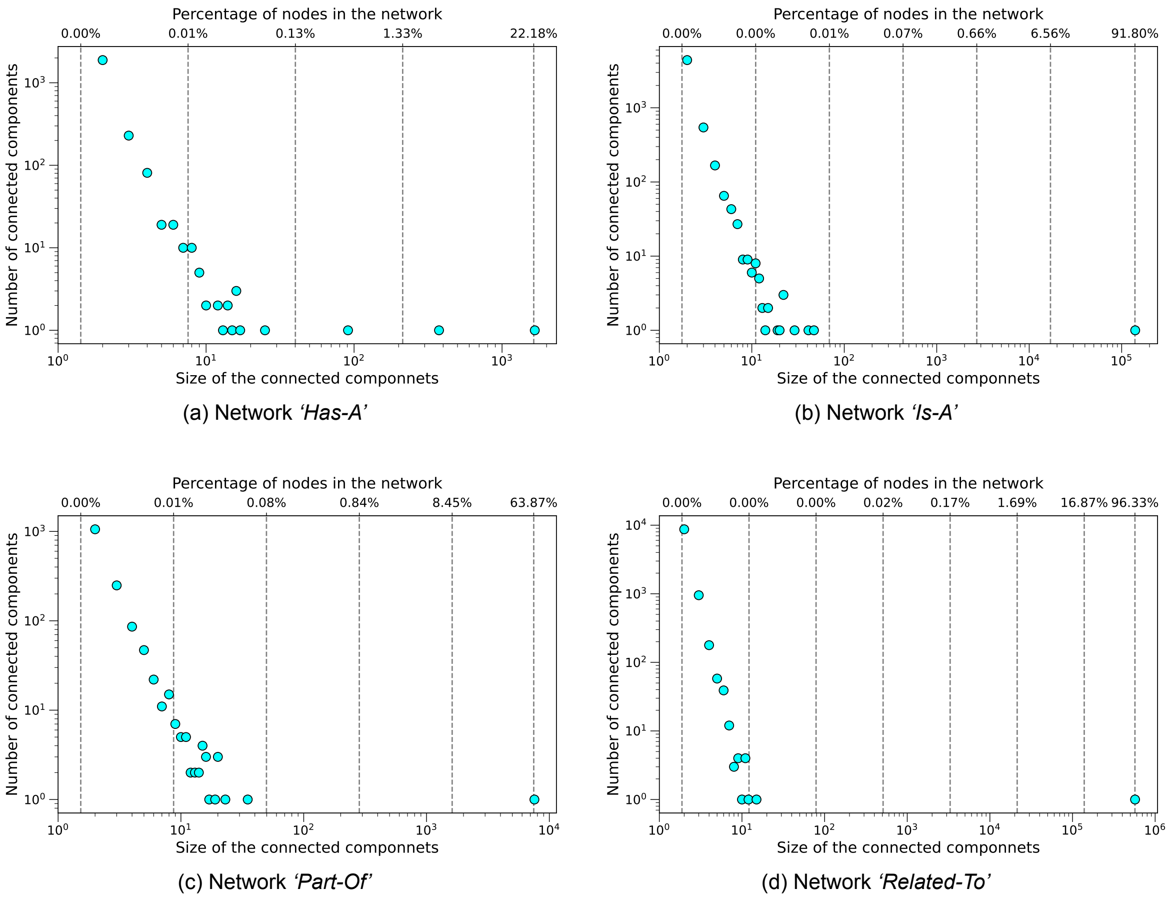
3.5. Connectedness

We first investigate the connectedness of the semantic networks. We measure the connectedness of a network by the size of the largest connected component and the size distribution of all connected components. Table 3.4 lists the sizes of the largest connected components and the percentage of nodes in corresponding networks. The same statistics are computed for the rewired semantic networks for comparison. Judging from the percentages of nodes in the largest connected component, all seven semantic networks are not fully connected. Networks 'Is-A', 'Related-To' and 'Union' are almost connected, since their largest connected components contain over 90% of nodes. As for the other networks, (*i.e.*, 'Has-A', 'Part-Of', 'Antonym' and 'Synonym'), they are widely disconnected.

Most rewired networks show more connectedness than corresponding real networks, especially for 'Antonym' and 'Synonym'. *In other words, the majority of semantic networks are less connected than expected by chance.* As for network 'Related-To' and 'Union', the percentage of the largest connected component remains almost unchanged. However, the 'Is-A' network is more connected than expected by chance.

Network	Size of full network	Size of LCC	Percentage
Has-A *Has-A (rewired)*	7,503	1,664 2,416 ± 35	22.18% (32.20 ± 0.47)%
Is-A *Is-A (rewired)*	152,538	140,024 127,258 ± 73	91.80% (83.43 ± 0.05)%
Part-Of *Part-Of (rewired)*	11,839	7,562 7,993 ± 53	63.87% (67.51 ± 0.45)%
Related-To *Related-To (rewired)*	592,816	571,079 570,012 ± 116	96.33% (96.15 ± 0.02)%
Union *Union (rewired)*	677,426	650,079 650,474 ± 182	95.96% (95.77 ± 0.03)%
Antonym *Antonym (rewired)*	16,867	5,912 8,845 ± 59	35.05% (52.44 ± 0.35)%
Synonym *Synonym (rewired)*	166,922	53,279 103,466 ± 142	31.92% (61.98 ± 0.09)%

Table 3.4: Size of Largest Connected Component (LCC) of the seven networks compare to their full and rewired networks. The size of LCC of each rewired network is the average of results obtained from 10 times rewiring with standard deviation.



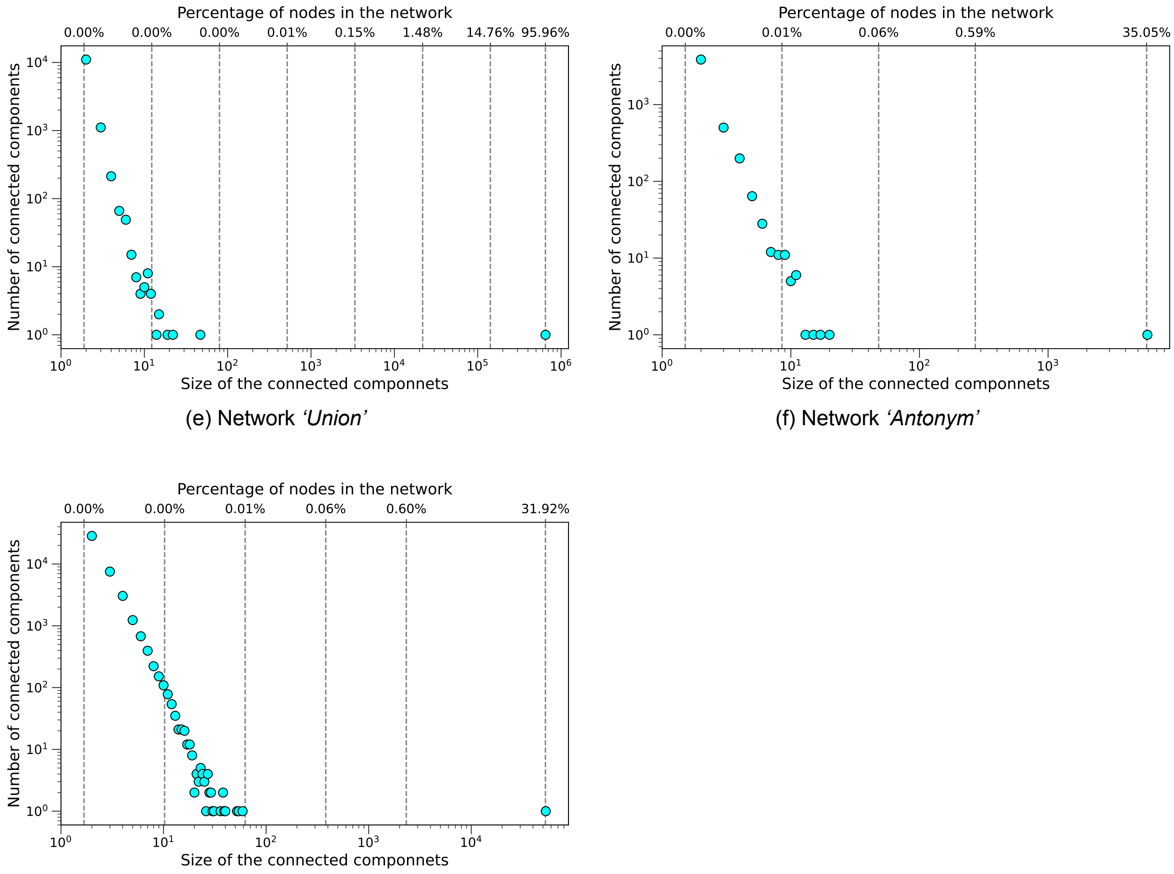
(a) Network *'Has-A'*

(b) Network *'Is-A'*

(c) Network *'Part-Of'*

(d) Network *'Related-To'*

Figure 3.2: Size distributions of connected components of the seven English semantic networks. The dashed lines indicate the percentage of nodes in connected components.



(e) Network *'Union'*

(f) Network *'Antonym'*

(g) Network *'Synonym'*

Figure 3.2: Size distributions of connected components of the seven English semantic networks. The dashed lines indicate the percentage of nodes in connected components (cont.)

Apart from the largest connected component, we are also interested in the size distribution of connected components in our semantic networks. Thus, we compute all connected components for each network and count the occurrence of different sizes of connected components. The results are presented in Fig. 3.2. *Overall, almost every network has a large connected component that is several orders of magnitude larger than the rest of the connected components, except for network 'Has-A', which has multiple larger connected components.* Network 'Has-A' is more fragmented. It has three relatively larger connected components, where the node with the largest degree is not in the largest connected component but in the second largest one. We checked these three connected components and discovered that their nodes have distinct themes. The component with the largest degree node contains all kinds of disease names. We believe that the fragmentation is caused by the manual and automatic creation of the dataset.

From now on, we restrict all further semantic network analyses to the largest connected components of all networks, unless otherwise mentioned.

3.6. Degree Distributions

The degree distribution captures the structure of a network. A common way to visualize the degree distribution is through a histogram. After obtaining the degree of all nodes of a network, we plot the probability $\Pr[D = k]$ against discrete degree k in a histogram using a bin width of 1. As shown in the log-log plots in the first column of Fig. 3.3, for seven networks, the probability $\Pr[D = k]$ decays in an almost straight line for larger values of k, followed by a fat tail. The linear dependency between $\log(\Pr[D = k])$ and $\log(k)$ in the tail confirms that *all semantic networks have power-law degree distributions*, which can be characterized by

$$\Pr[D = k] \sim k^{-\gamma}, \tag{3.1}$$

where γ is the power-law exponent.

Next, we estimate the power-law exponent γ of each network. This step is done by estimating the slope of a degree distribution using linear regression. To obtain better estimation, we implement *logarithmic binning* (see Appendix A) to suppress the noise at larger values of the degree k (the fat tail). The second column of Fig. 3.3 depicts the results of degree distributions using logarithmic binning and linear regression. The estimated power-law exponents γ of the seven networks are listed in Table 3.5.

Network	***Has-A***	***Is-A***	***Part-Of***	***Related-To***	***Union***	***Antonym***	***Synonym***
γ	2.3	2.3	2.4	2.4	2.4	2.5	3.7

Table 3.5: The power-law exponents γ of the seven English semantic networks.

Discussion The power-law exponent γ of most semantic networks lies between 2 and 3, except for the 'Synonym' network ($\gamma = 3.7$). Because of $2 < \gamma < 3$, we expect these networks to have a finite average degree but a very large variance, which can be explained by the n^{th} moment of the degree distribution [33, 34]

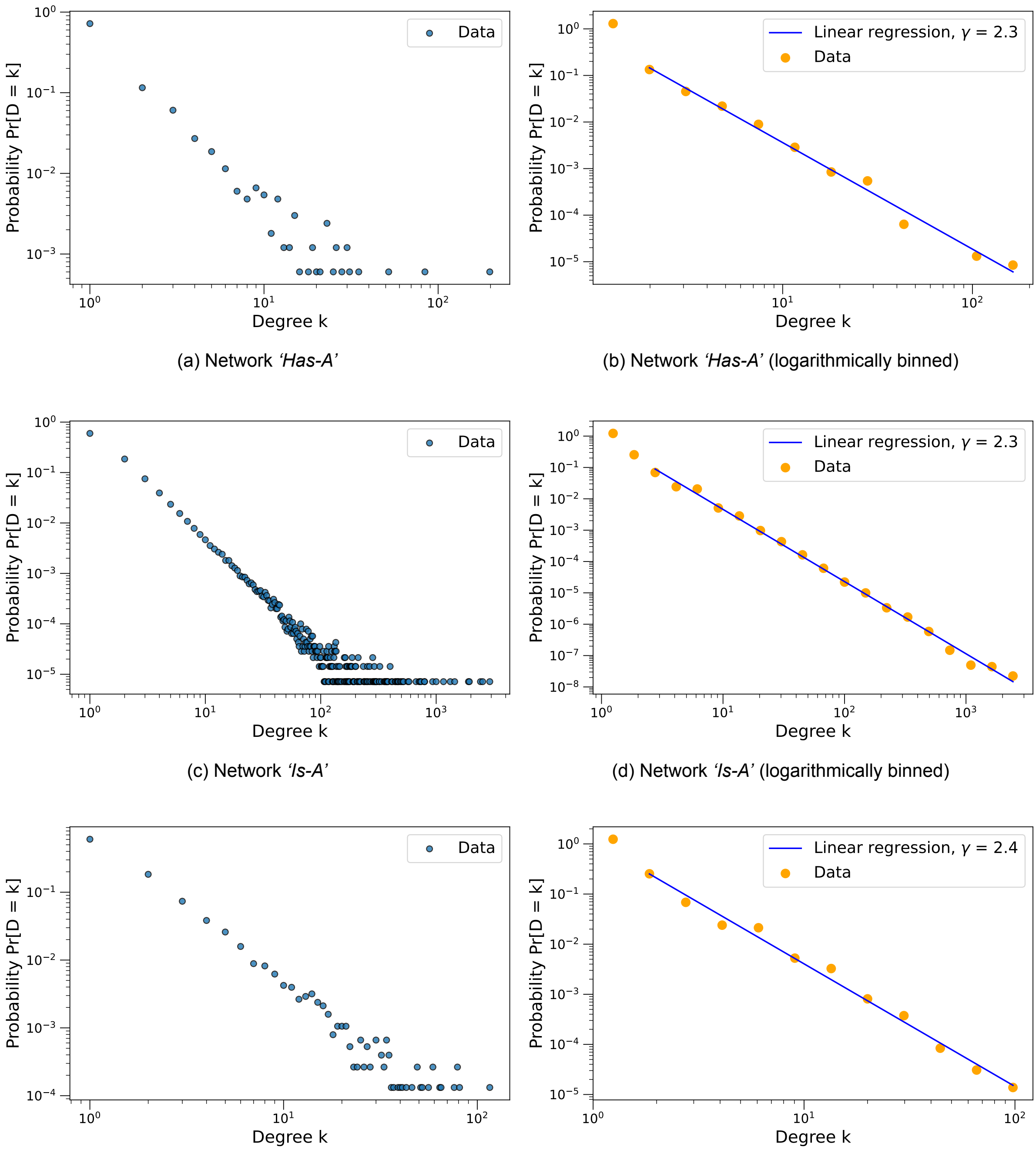
$$E[D^n] = \sum_{k=1}^{\infty} k^n \Pr[D = k]. \tag{3.2}$$

The first moment ($n = 1$) is the average degree $E[D]$. The second moment $E[D^2]$ is related to the variance $\sigma^2 = E[D^2] - E[D]^2$.

For a network that has a degree distribution with a power-law exponent $\gamma \in (2, 3)$, the first moment $E[D]$ is finite but the second moment $E[D^2]$ is infinite. Networks with this property are known as scale-free networks [35, 55]. The average degree of these network is not representative, because the variance is very large. The name 'scale-free' indicates that there is no characteristic scale for networks with a power-law degree distribution. In scale-free networks, nodes have widely different degrees, there are many nodes with small degree and a few nodes with very large degree.

We find that *most semantic networks are scale-free* networks. This coincides with the findings in most literature. That is semantic networks are highly heterogeneous [53, 61]. There are many specific or unique words that can be paired with only a few other words, but there are also some general words that can be matched with almost anything. We relate the generality of a word with its degree, the more general a word is, the larger its degree. Examples of general words are 'plant', 'water', 'person', 'time', etc. Whereas, words like ‘neotectonic’, ‘ungraced’, ‘cofinance’ and ‘informatically’ are much less general.

Nevertheless, we should not ignore that the power-law degree distribution has been observed in an abundance of networks, ranging from social [62], biological [63] to communication networks [64]. Our results show that semantic networks, like many other types of networks, have power-law distributions in their degree.



(a) Network *‘Has-A’*

(b) Network *‘Has-A’* (logarithmically binned)

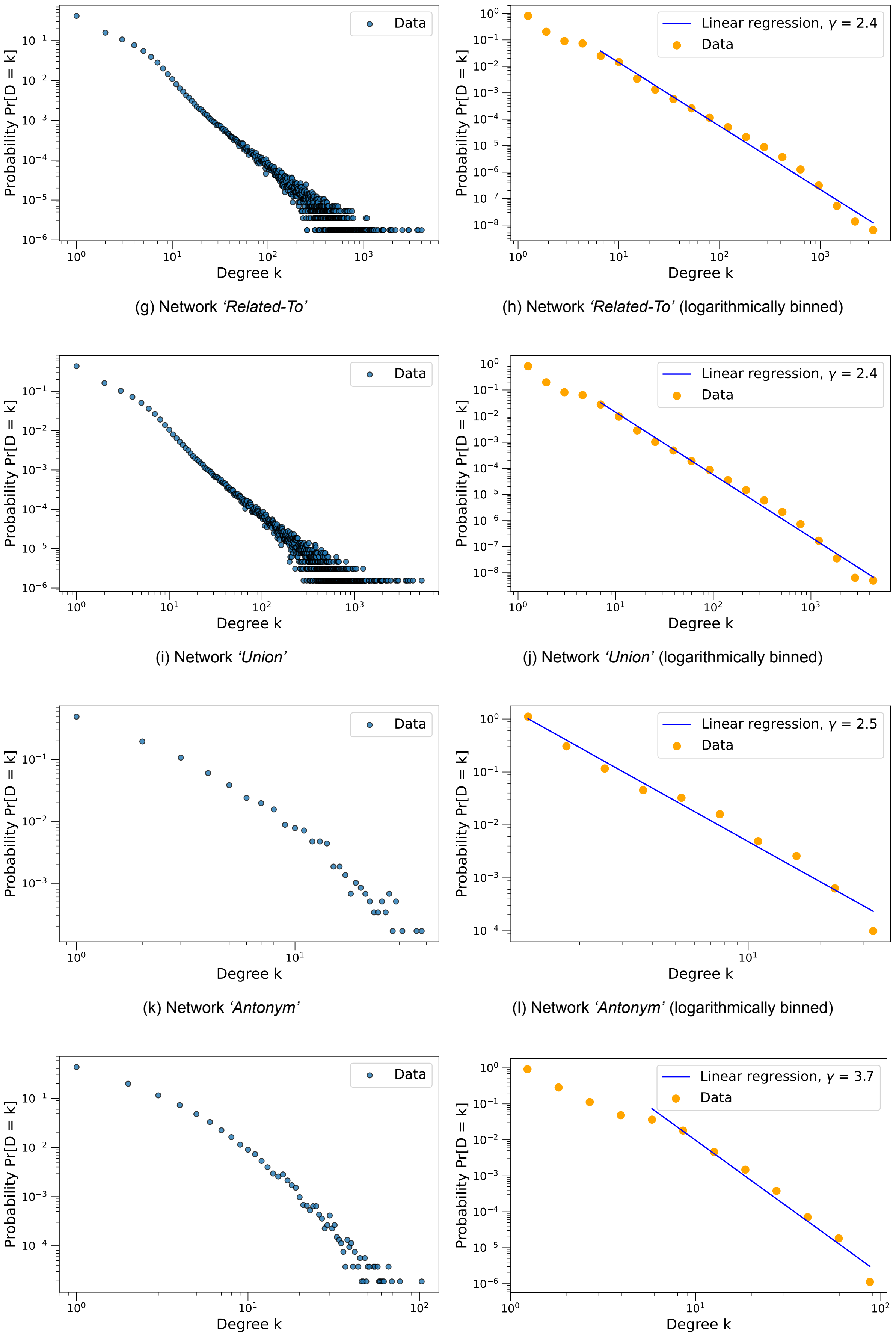
(c) Network *‘Is-A’*

(d) Network *‘Is-A’* (logarithmically binned)

(e) Network *‘Part-Of’*

(f) Network *‘Part-Of’* (logarithmically binned)

Figure 3.3: Degree distributions of seven English semantic networks and power-law exponent estimation over logarithmically binned degree distribution. The regression fitting does not always start from the first data point. Because we focus on the linear part at the larger values of the degree k, we inspect the degree distribution of each network and exclude the non-linear part.



(g) Network *'Related-To'*

(h) Network *'Related-To'* (logarithmically binned)

(i) Network *'Union'*

(j) Network *'Union'* (logarithmically binned)

(k) Network *'Antonym'*

(l) Network *'Antonym'* (logarithmically binned)

(m) Network *'Synonym'*

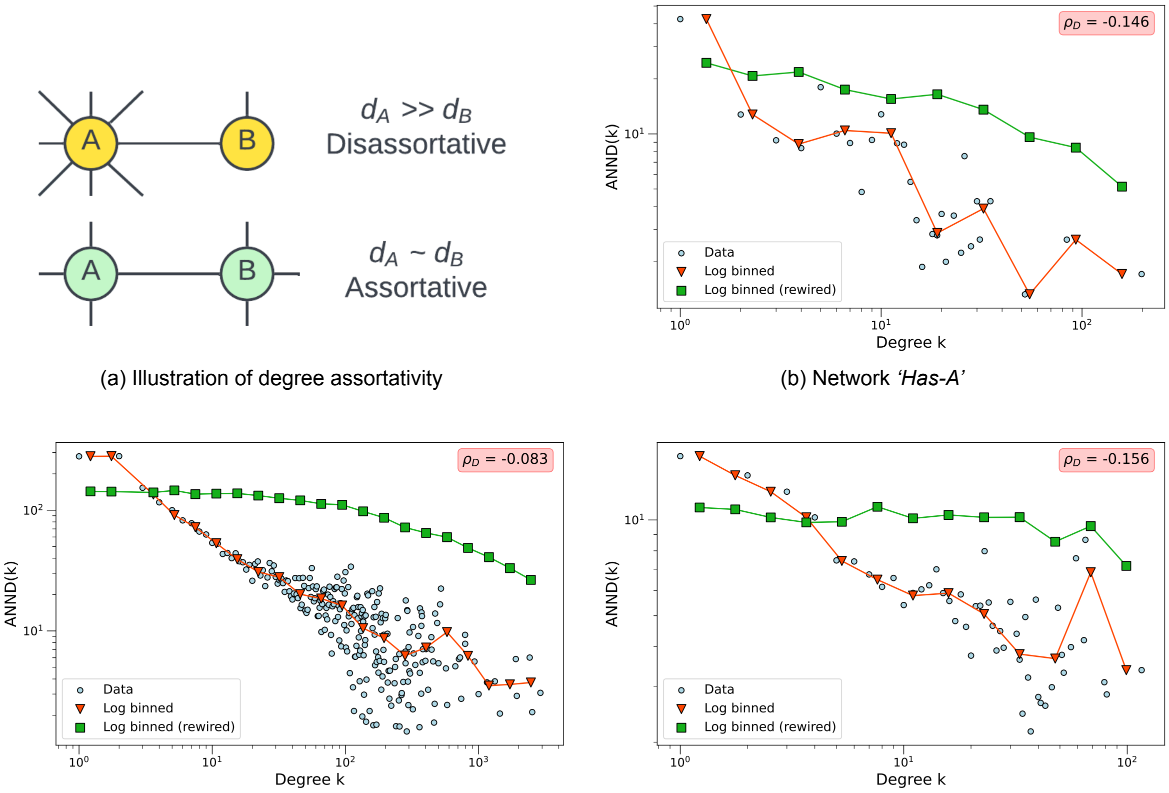
(n) Network *'Synonym'* (logarithmically binned)

Figure 3.3: Degree distributions of seven English semantic networks and power-law exponent estimation over logarithmically binned degree distribution. The regression fitting does not always start from the first data point. Because we focus on the linear part at the larger values of the degree k, we inspect the degree distribution of each network and exclude the non-linear part (cont.).

3.7. Degree Assortativity

After inspecting the node connectivity, we would like to learn what are the mixing patterns in these networks. Do nodes tend to connect to nodes with similar degree? Or alternatively, do larger-degree nodes tend to connect with small-degree nodes? This property of networks is known as degree assortativity. One can distinguish the latter type of networks as disassortative, while the former type of networks are assortative. There have been established a number of measures to quantify the degree assortativity. One of them is the degree correlation coefficient and another one is the Average Nearest Neighbor Degree (ANND), both of which we defined in Section 2.2.2.

We plot the average nearest neighbor degree as a function of the degree k. Fig. 3.4 depicts the function ANND(k) together with the degree correlation coefficient ρ_D. Meanwhile, for every network, we calculate the ANND(k) of its rewired network for comparison. Randomized networks with preserved degree distribution have no degree-degree correlation. As a result, the function ANND(k) dose not vary with k. Therefore, we use these randomized networks as a reference to see the ANND values we could expect when the links are distributed at random.



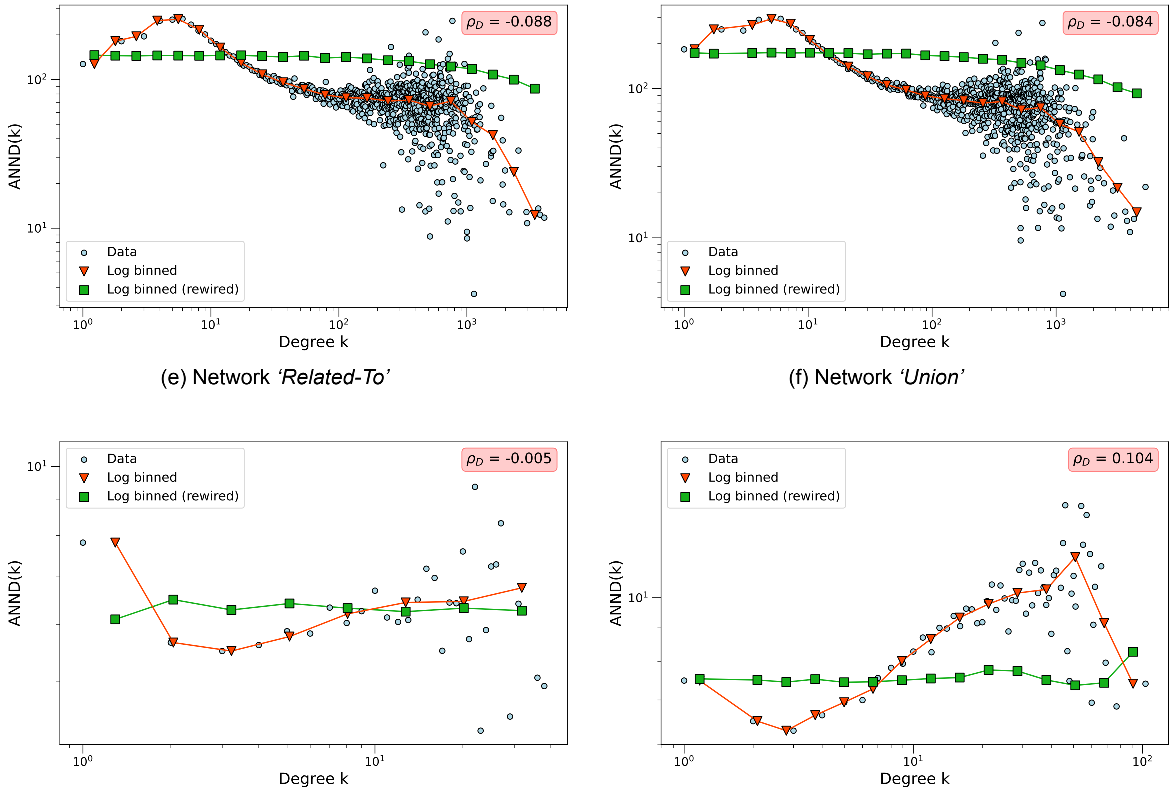
(a) Illustration of degree assortativity

(b) Network *'Has-A'*

(c) Network *'Is-A'*

(d) Network *'Part-Of'*

Figure 3.4: Degree assortativity of semantic networks. (a) Examples of disassortative and assortative mixing; (b-h) Average nearest neighbor degree (ANND) as a function of degree k and degree correlation coefficient ρ_D of seven English semantic networks. Data points in light blue are the average ANND of nodes with degree k in a network, red triangles represent the data after logarithmic binning, and green squares are the average ANND of nodes with degree k in the rewired network. Note logarithmic binning is applied to reduce the noise and better visualize the data.



(e) Network *'Related-To'*

(f) Network *'Union'*

(g) Network *'Antonym'*

(h) Network *'Synonym'*

Figure 3.4: Degree assortativity of semantic networks. (a) Examples of disassortative and assortative mixing; (b-h) Average nearest neighbor degree (ANND) as a function of degree k and degree correlation coefficient ρ_D of seven English semantic networks. Data points in light blue are the average ANND of nodes with degree k in a network, red triangles represent the data after logarithmic binning, and green squares are the average ANND of nodes with degree k in the rewired network. Note logarithmic binning is applied to reduce the noise and better visualize the data (cont.).

Discussion Fig. 3.4 illustrates that *most semantic networks are disassortative* as ANND(k) is a decreasing function over degree k. These networks are 'Has-A', 'Part-Of', 'Is-A', 'Related-To' and 'Union'. The negative degree correlation coefficients also validate the disassortativity. In disassortative networks, nodes with larger degree (general words) tend to connect to nodes with smaller degree (less general words). This is not surprising as when we use these relations, 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(k) increases in the degree k. This indicates that large-degree nodes (general words) associate with nodes that have similar degree (words with the same generality). The same applies for network 'Antonym', though the degree correlation is not very visible, we still see a slight upward trend in the curve of ANND(k). This is also reflected in the small correlation coefficient $\rho_D = -0.005$.

Note that the function ANND(k) of a rewired network is not degree-dependent anymore (see the green curves in Fig. 3.4). For example, the curve is almost flat for 'Synonym' and 'Related-To'. At the larger degree k, the curve may drop slightly. This induced disassortativity is caused by large-degree nodes having not enough neighbors to connect to.

3.8. Clustering Coefficient

In networks such as social networks, the neighbors of a node tend to be connected as well. This tendency is known as clustering [65]. If a person has a group of friends, there is a high chance that these friends also know each other. In such networks, there are lots of triangular connections. What is the clustering in semantic networks? In this section, we investigate this property by measuring the clustering coefficient c_G. The clustering coefficient c_G is a measure of how closely nodes in a network cluster together.

We calculate the average local clustering coefficient $c_G(k)$ of nodes with degree k for each network. Fig. 3.5 demonstrates the results. In addition, we calculate the $c_G(k)$ for rewired networks as a guideline for comparison with the semantic networks. To compare the clustering of semantic networks and a completely random network, we calculate the average clustering coefficient of an Erdős–Rényi (ER) random graph with the same number of nodes N and links L, indicated by the yellow line in Fig. 3.5. The average clustering coefficient of an ER random graph $G(N, L)$ is simply $p = E[D]/(N-1)$, the proof is given in [39].

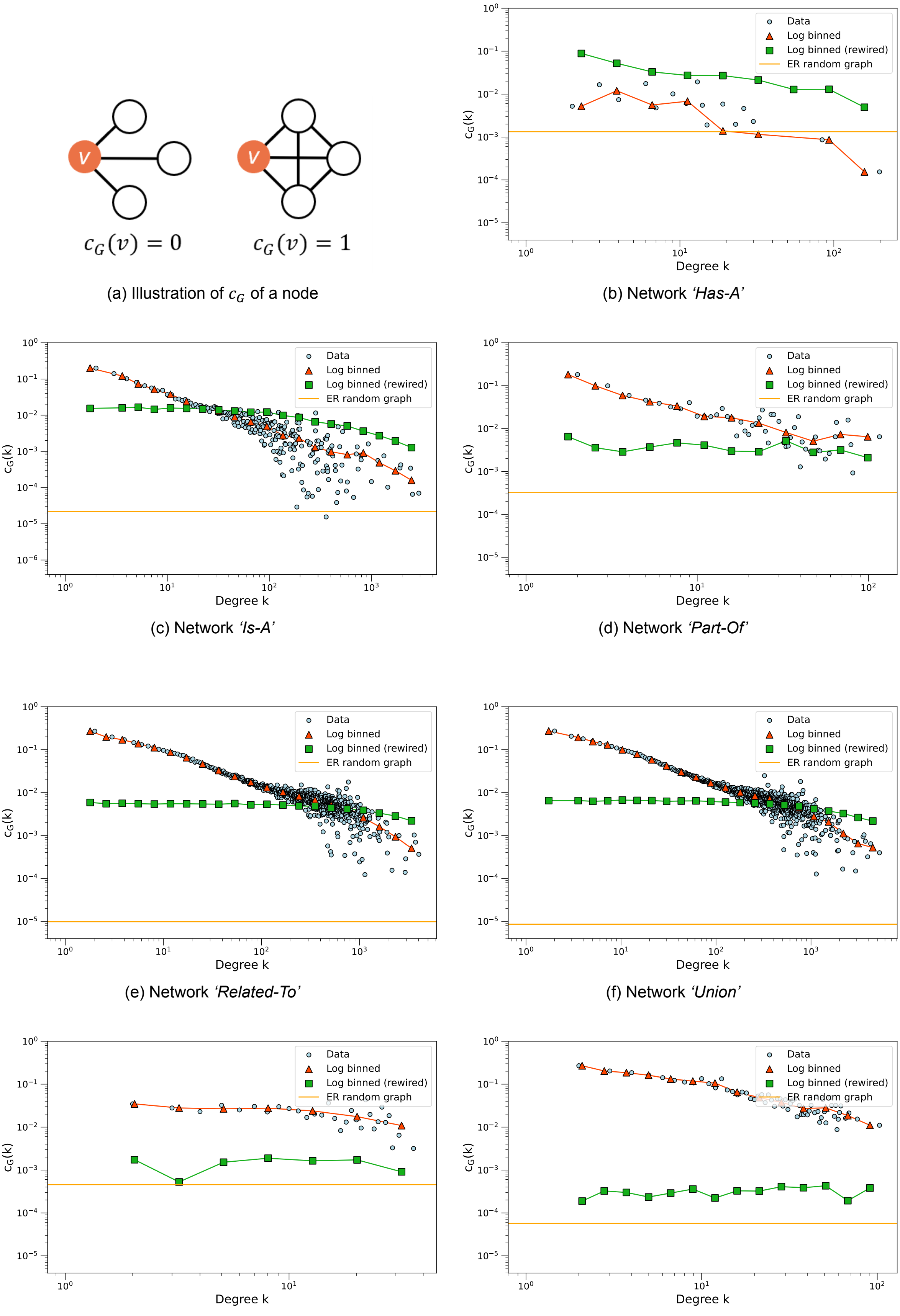
Discussion Fig. 3.5 shows that all semantic networks have much larger average clustering coefficients c_G than the ER random graph, except for 'Has-A'. Because in a random network with a large number of nodes N and relatively small number of links L (sparse), the probability of three nodes forming a triangle is very small.

Moreover, there are some semantic networks that have substantially larger clustering coefficients than their randomized versions (rewired), *i.e.*, 'Part-Of', 'Antonym' and 'Synonym'. There are more triangles in these networks than expected by chance.

On the other hand, network 'Has-A' has lower clustering coefficients $c_G(k)$ than the randomized network. Our hypothesis is that the 'Has-A' network is organized differently from the other networks, as there are fewer triangles.

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

We discover that networks with different link types show different degrees of clustering. This encourages us to to further inspect the organizing principles of these semantic networks, which we will discuss explicitly in Chapter 5.



(a) Illustration of c_G of a node

(b) Network *'Has-A'*

(c) Network *'Is-A'*

(d) Network *'Part-Of'*

(e) Network *'Related-To'*

(f) Network *'Union'*

(g) Network *'Antonym'*

(h) Network *'Synonym'*

Figure 3.5: Clustering coefficient of semantic networks. (a) Toy examples of clustering coefficient of a node v; (b-h) The average local clustering coefficient $c_G(k)$ of nodes with degree k of seven English semantic networks (in light blue data points). Red triangles represent data after logarithmic binning, and green squares are the average clustering coefficient of nodes with degree k (logarithmically binned) in the rewired networks. The yellow horizontal line indicates the implied c_G of an Erdős–Rényi (ER) random graph with the same number of nodes and links.

3.9. Statistics of Semantic Networks

We calculated the overall descriptive statistics of semantic networks, the number of nodes N, number of links L, the maximal degree d_{max}, the average degree $E[D]$, average nearest neighbor degree (ANND), the average clustering coefficient c_G and the graph transitivity $\check{c}_G$. Besides, we rewire and reconstruct all semantic networks using the methods described in Section 2.3. The same statistics are calculated for the rewired and reconstructed networks. Additionally, we also include the power-law exponents. The results are summarized in Table 3.6.

For networks obtained by degree-preserving rewiring and reconstruction, only the ANND, the average clustering coefficient c_G and graph transitivity $\check{c}_G$ change. The statistics of both random networks, rewired and reconstructed, are quite similar. The average nearest neighbor degree ANND becomes smaller for all randomized semantic networks except for 'Synonym'.

As for the average clustering coefficient, all networks except 'Has-A' have remarkably larger c_G (at least by an order of magnitude) than the randomized networks. Because in random networks links are randomly distributed, there are fewer triangles. On the contrary, randomized networks of 'Has-A' exhibit more than seven times larger clustering coefficients than its original network.

The graph transitivity provides the ratio of the number of closed triangles in a network relative to the total possible triples. We see that the graph transitivity and the average clustering coefficient are different for every network. All networks have smaller graph transitivity $\check{c}_G$ than the average clustering coefficient c_G, except for 'Antonym'. And the c_G and $\check{c}_G$ show big discrepancy in 'Related-To' and 'Union'. This discrepancy may be related to certain motifs in a network as Estrada [66] pointed out. Due to limited time, we leave the investigation of the discrepancy in graph transitivity and average clustering coefficient for future work.

Network	***Has-A***	***Is-A***	***Part-Of***	***Related-To***	***Union***	***Antonym***	***Synonym***
N	1,664	140,024	7,562	571,079	650,079	5,912	53,279
L	1,842	213,319	9,212	1,598,548	1,804,666	7,986	80,668
d_{max}	198	2913	116	4025	5263	38	103
$E[D]$	2.21	3.05	2.44	5.6	5.55	2.7	3.03
ANND	33.6	242	14.1	170	219	6.77	7.13
ANND rewired	23.3	142	10.8	145	173	6.25	7.51
ANND reconstructed	22.8	137	10.6	144	171	6.26	7.54
c_G	2.17×10^{-3}	5.66×10^{-2}	4.61×10^{-2}	1.02×10^{-1}	1.04×10^{-1}	1.50×10^{-2}	1.13×10^{-1}
c_G *rewired*	1.83×10^{-2}	6.26×10^{-3}	1.95×10^{-3}	3.26×10^{-3}	3.68×10^{-3}	7.26×10^{-4}	1.48×10^{-4}
c_G *reconstructed*	1.56×10^{-2}	7.11×10^{-3}	1.29×10^{-3}	3.33×10^{-3}	3.86×10^{-3}	7.42×10^{-4}	9.68×10^{-5}
$\check{c}_G$	1.16×10^{-3}	2.20×10^{-3}	1.79×10^{-2}	8.01×10^{-3}	7.22×10^{-3}	2.18×10^{-2}	9.07×10^{-2}
$\check{c}_G$ *rewired*	1.28×10^{-2}	3.80×10^{-3}	3.53×10^{-3}	4.09×10^{-3}	4.35×10^{-3}	1.56×10^{-3}	3.14×10^{-4}
$\check{c}_G$ *reconstructed*	1.34×10^{-2}	5.30×10^{-3}	3.98×10^{-3}	4.61×10^{-3}	5.13×10^{-3}	1.98×10^{-3}	2.45×10^{-4}
γ	2.3	2.3	2.4	2.4	2.4	2.5	3.7

Table 3.6: Statistics of largest connected component of seven English semantic networks extracted from ConceptNet.

In summary, we find universalities of semantic networks across *degree distribution*, *degree assortativity*, *clustering*, *sparsity* and *connectedness*.

- All semantic networks have power-law degree distribution and most of them are scale-free networks.

- There are two types of degree mixing patterns in semantic networks, assortative and disassortative.

- Most networks have higher average clustering coefficients than expected by chance, except for one network, ‘Has-A’, which shows lower clustering.

- All semantic networks have high sparsity.

- Most networks have a single connected component with the majority of nodes, except for network ‘Has-A’, which is more fragmented.

4

Semantic Networks in Different Languages

We have investigated the general topology and found universal characteristics of seven English semantic networks. Nevertheless, there are thousands of other languages in the world. Do semantic networks in other languages possess the same topological properties as in English? In this Chapter, we zoom out from the English semantic networks and consider many other languages in ConceptNet. Based on the vocabulary size and sources of knowledge, we chose 10 languages besides English. They are French, Italian, German, Spanish, Russian, Portuguese, Dutch, Japanese, Finnish and Chinese. We extract these semantic networks using the same procedure explained in Section 3.3.

Foremost, we classify the 11 languages into several language families. Then we calculate the basic statistics for seven semantic networks (seven link types) for eleven languages. To explore whether there exist special patterns in semantic networks in different languages, we compare several topological properties based on language families. At last, we present an interesting phenomenon, language inflection, that we observe in the degree distribution of the 'Related-To' networks.

4.1. Language Families

In linguistics, languages are grouped into multiple categories according to different rules. There are two kinds of language classifications: genetic and typological.

Genetic classification, also known as genealogical classification, assorts languages according to their level of diachronic relatedness [67]. In other words, languages are categorized into the same family if they evolved from the same root language. Take one of the world's primary language families, Indo-European, for example. This family has several branches such as Germanic, Balto-Slavic and Italic [68]. Moreover, Germanic languages include English, German, Danish, etc.

Typological classification classifies languages based on their structural features. One popular typological classification distinguishes isolating, agglutinating and inflecting languages. It groups languages in accordance with the morphological formation of words. A morph (or morpheme) is the basic unit of a word [69] such as stems and affixes. For instance, the word 'undoubtedly' consists of three morphs: 'un-', 'doubted' and '-ly'. In an *isolating* language, each word contains only a single morph [67]. One particular example of a highly isolating language is Chinese. On the contrary, words in an *agglutinating* language can be divided into morphs with distinctive grammatical categories such as tense, person and gender. But in an *inflecting* language, there is no exact match between morphs and grammatical categories [67]. A word changes its form depending on different grammars. Most Indo-European languages belong to this inflecting family.

Based on these two classifications, we divide the 11 selected languages (including English) into several language families. Table 4.1 specifies the sub-families of typological and genetic classifications respectively. Typologically, most of the languages that we study (8 out of 11) belong to the inflecting family. This classification is more general since it only has three categories. While genetic classification identifies more sub-families and distributes the eleven languages more evenly. Therefore, we adopt six genetically classified language families in our semantic network analyses. Nevertheless, we will make use of the typological classification as a reference.

Typological Genetic	**Inflecting**	**Isolating**	**Agglutinating**
Italic	French, Italian Spanish, Portuguese		
Germanic	English, Dutch German		
Balto-Slavic	Russian		
Transeurasian			Japanese
Sino-Tibetan		Chinese	
Uralic			Finnish

Table 4.1: Sub-families of typological and genetic classifications of eleven languages chosen from ConceptNet.

4.2. Overview of Semantic Networks in Eleven Languages

For every language, we construct seven undirected semantic networks with separated link types: 'Has-A', 'Part-Of', 'Is-A', 'Related-To', 'Union', 'Antonym' and 'Synonym'. Due to a data availability limitation, only three languages have the 'Has-A' relation. For languages without the 'Has-A' relation, the 'Union' network is just the union of three link types: 'Part-Of', 'Is-A' and 'Related-To'. In this section, we give an overview of the size and average degree of the semantic networks. Again, we restrict our study to the largest connected component of these networks.

Table 4.2 shows the number of nodes of each semantic network in eleven languages. A blank element in the table indicates that the network does not exist, *i.e.*, a relation is unavailable in a language. Regarding the number of nodes in these networks, 'Related-To' and 'Union' are generally the largest networks, in which the French 'Union' network is the largest. At the same time, there are many small networks with size $N < 100$, particularly in 'Part-Of' and 'Synonym' networks. In the following sections, we will focus on the larger networks.

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
English	1,664	140,024	7,562	571,079	650,079	5,912	53,279
French		17,519	2,832	1,289,083	1,296,622	1,361	20,144
Italian		2,663	9	36,295	46,468	13	1,580
German		113,301	5	100,737	172,147	187	43,072
Spanish		255	11	12,094	22,861	15	3,491
Russian		557	3	20,268	25,887	12	1,148
Portuguese		3,341	15	5,929	11,426	17	6,421
Dutch		191	53	303	1,418	111	11,964
Japanese	38	40,256	7,230	7,200	43,286	20	230
Finnish		76	12	4,483	6,958	24	1,569
Chinese	6,355	10,073	3,417	3,163	17,128	4	17

Table 4.2: Number of nodes N of semantic networks in the eleven languages extracted from ConceptNet. A blank element indicates the corresponding network is unavailable. The 'Union' network is the union of four networks ('Has-A', 'Is-A', 'Part-Of' and 'Related-To'). Because we display largest connected component sizes, for some 'Union' networks, the number of nodes exceeds the sum of the sizes of its four constituent networks.

Similar to English semantic networks, we observe that most networks (with more than 100 nodes) in other languages are sparse. Table 4.3 lists the average degree $E[D]$ of all semantic networks. We can see that all networks have an average degree between 1 and 6. Consider the Dutch 'Is-A' network for example, a node has about 5 connections on average, which is only 2.45% of 191 nodes in the whole network.

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
English	2.21	3.05	2.44	5.6	5.55	2.7	3.03
French		2.64	2.51	3.44	3.46	2.45	2.81
Italian		2.86	2.89	2.2	2.27	1.85	2.54
German		2.75	1.6	4.77	4.53	2.16	3.57
Spanish		2.45	2.73	2.13	2.13	1.87	2.57
Russian		2.23	1.33	4.14	3.88	1.83	2.26
Portuguese		2.24	2.67	2.49	2.65	2	2.84
Dutch		4.68	4.98	2.3	2.69	2.11	3.53
Japanese	2.89	4.42	4.11	4.34	4.79	2	2.73
Finnish		1.97	1.83	2.3	2.26	1.92	2.24
Chinese	3.58	3.02	3.36	4.06	3.78	1.5	2.24

Table 4.3: Average degree $E[D]$ of semantic networks in the eleven languages extracted from ConceptNet. A blank element indicates the corresponding network is unavailable.

In the subsequent sections, we compare three principal topological properties of semantic networks in eleven languages in general. Specifically, degree distribution, degree correlation coefficient and clustering. Complete statistics of topological properties of the full networks and corresponding largest connected components in every language can be found in Appendix B.

4.3. Degree Distribution

First, we look at the degree distribution of each of the semantic networks. We estimate the power-law exponents for networks with size $N > 1000$. The reason we leave out networks with fewer than 1000 nodes is that we need sufficient data to estimate the power-law exponent γ. Appendix D includes the plots of the degree distributions of all semantic networks (and the logarithmically binned version if there is a power-law).

Table 4.4 lists the estimated power-law exponent γ for each semantic network in the eleven languages. We consider a network to not have a power-law degree distribution if the distribution (log-log scale) clearly deviates from a straight line at larger values of degree k.

Additionally, we are interested in the power-law degree distribution with $2 < \gamma < 3$. Thus, we plot the estimated power-law exponent γ of semantic networks for better comparison (Fig. 4.1).

Network	***Has-A***	***Is-A***	***Part-Of***	***Related-To***	***Union***	***Antonym***	***Synonym***
English	2.3	2.3	2.4	2.4	2.4	2.7	3.7
French		2.4	2.3	X	X	2.7	3.1
Italian		2.3	None	2.6	2.6	None	3.7
German		2.5	None	2.6	2.5	None	3.1
Spanish		None	None	X	X	None	3.0
Russian		None	None	3.6	3.5	None	3.0
Portuguese		2.6	None	2.4	2.5	None	4.4
Dutch		None	None	None	2.2	None	4.8
Japanese	None	2.4	2.3	2.2	2.3	None	None
Finnish		None	None	X	X	None	4.2
Chinese	2.5	2.3	2.7	1.9	2.3	None	None

Table 4.4: Power-law exponent γ of semantic networks in different languages. The γ is shown as 'None' for networks with size $N < 1000$. A cross (X) represents that the degree distribution of that network is not a power-law.

Discussion Fig. 4.1 tells us that *most networks are scale-free*, with a few exceptions. Specifically, Chinese 'Related-To' has a $\gamma < 2$, Russian 'Related-To' and 'Union' both have power-law exponents larger than 3.

From Table 4.4, we notice that all 'Synonym' networks have a $\gamma \geq 3$. The reason for 'Synonym' networks to have such high power-law exponents is that their nodes have smaller degree compared to other networks. As a result, the slope of the degree distribution is steeper. This is not strange, since most words only have a certain amount of words that have similar meanings as them. It is quite difficult to find a word that has many synonyms.

Another interesting phenomenon is that the degree distributions of several 'Related-To' and 'Union' networks are not perfect power-laws, *i.e.*, networks for French, Spanish and Finnish. We discuss this phenomenon explicitly in Section 4.4.

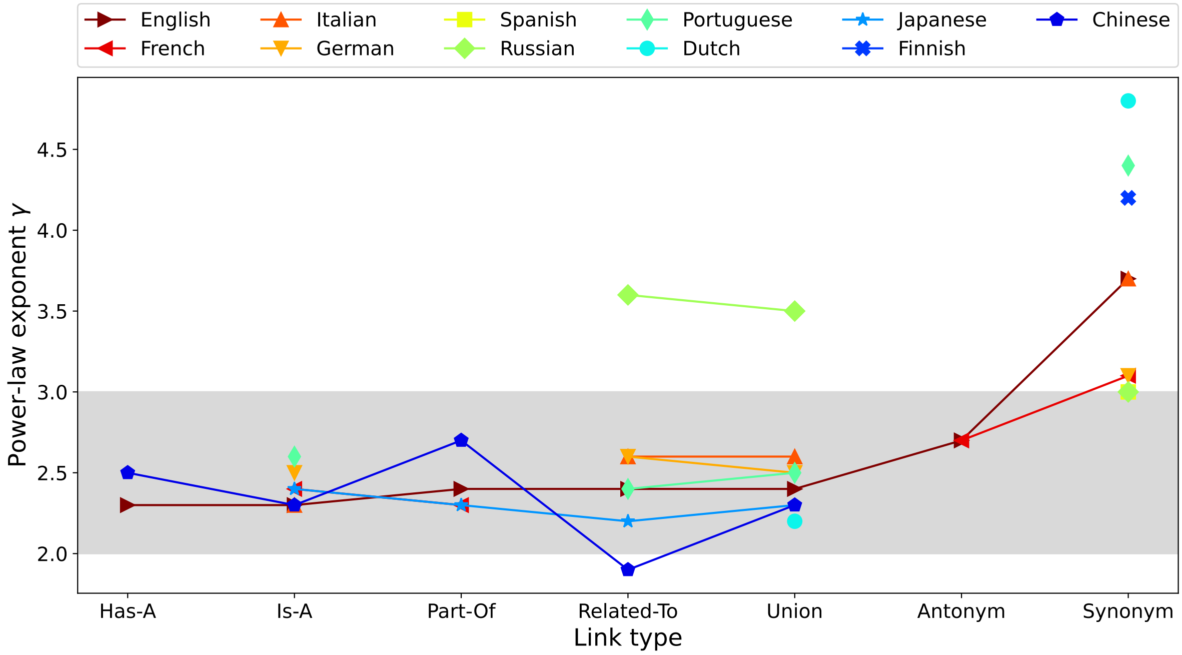


Figure 4.1: Power-law exponent γ of semantic networks in different languages. The range where $2 < \gamma < 3$ is shaded in grey.

4.4. Language Inflection

During the investigation of the degree distribution of networks 'Related-To' and 'Union', we notice some peculiar features. That is, for some languages, their degree distributions are not a perfect power-law. Instead, there are peaks in the distributions that lead to a deviation from a power-law. An example is the Spanish 'Related-To' network (see Fig. 4.2). We observe a peak in the tail of the distribution. Why do these nodes have such a high degree? This phenomenon encourages us to find its origin.

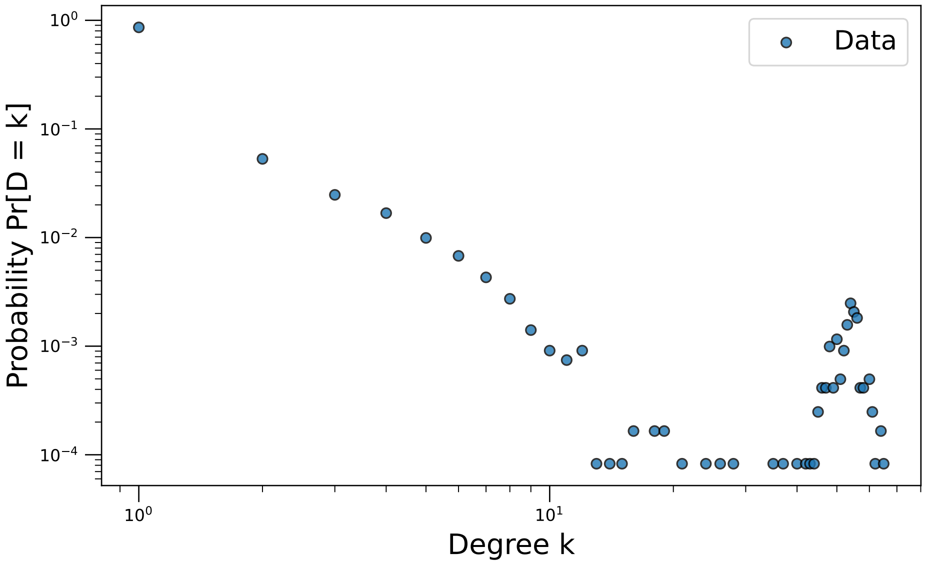


Figure 4.2: Degree distribution of the Spanish 'Related-To' network.

To really understand the phenomenon in the degree distribution, we look into the words in the peak and their neighboring words in the Spanish ‘Related-To’ network. Table 4.5 provides some examples of these peak words.

We find that most of the peak words are not only verbs, but also similar in the written forms. Hence, we suspect that these are grammatical inflections of different words. Our hypothesis is inspired by the knowledge that Spanish is a highly inflected language. Furthermore, we observe a similar anomaly in the degree distributions of French, Portuguese and Finnish ‘Related-To’ and ‘Union’ networks.

Peak word	English meaning	**Neighbors**
cenar	dine	cená, cenábamos, cenáculo, cenáis, cenáramos, cenáremos, ...
viajar	travel	viaja, viajaba, viajabais, viajaban, viajabas, viajad, viajado, ...
pasear	walk	pasea, paseaba, paseabais, paseaban, paseabas, pasead, ...
reparar	repair	repararais, repararan, repararas, reparareis, repararemos, ...
comparar	compare	comprar, comparaba, comparabais, comparaban, comparabas, ...

Table 4.5: Examples of words in the peak and their neighboring words in the Spanish ‘Related-To’ network.

Therefore, we decide to investigate whether this anomaly is caused by language inflection. From Table 4.2 we learn that network ‘Union’ is mostly composed of ‘Related-To’ in these four languages, thus, we restrict the analyses to the ‘Related-To’ networks.

4.4.1. Percentage of Word Types

In the previous section, we discover that words in the peak are mostly verbs for the Spanish ‘Related-To’ network. And similar anomaly in the degree distributions is observed in French, Portuguese and Finnish networks. To be concise, we refer to these words as *peak words* from now on. In this section, we want to know the significance of verbs in the peak words. To know the composition of these words better, we first compute the percentage of different word types in the peak words.

In ConceptNet, the word type of a node is provided in assertions. There are four types of words: verb, noun, adjective and adverb. However, the word type is not available for all nodes. Therefore, we only count the percentage based on nodes with known word types. We also calculate the percentage of four word types for the whole network (the largest connected component) for comparison. The percentages of verbs, nouns, adjectives and adverbs in both peak words and the whole ‘Related-To’ network are shown in Table 4.6. We can see that for all four languages, nouns and verbs are the major word types both in the LCCs and among the peak words.

In French, Spanish and Portuguese the major words are verbs. Furthermore, the percentage of verbs in the peak is larger than in the LCC. For example, 100% of Portuguese peak words are verbs.

However, for Finnish, the majority of peak words are nouns. Moreover, the percentage of nouns in the peak is larger than in the LCC.

Percentage (%)	French		Spanish		Portuguese		Finnish	
	LCC	Peak	LCC	Peak	LCC	Peak	LCC	Peak
Words with types	98.71	98.66	92.72	77.84	67.60	60.00	81.37	64.13
Verb	68.90	89.97	87.62	98.44	32.56	100.00	11.40	11.36
Noun	19.21	7.14	9.20	1.56	51.96	0	77.96	84.09
Adjective	11.53	2.75	2.89	0	14.60	0	7.17	4.55
Adverb	0.36	0.15	0.29	0	0.88	0	3.47	0

Table 4.6: Percentages of four word types among peak words and in the Largest Connected Component (LCC) of network 'Related-To' in four inflecting languages.

The high percentage in one type of words in peak words make us wonder if there is indeed special grammatical structure involved. So we are curious about the neighbors of each peak word, *i.e.* the type of words of these neighbors. Since verbs and nouns are the major types, we focus on these two types in the following analyses.

To obtain the percentage of verbs and nouns of all neighbors of peak words, we first compute the percentage in the neighbors of each peak word. Then we take the mean of all percentages. The mean percentage and standard deviation are presented in Table 4.7.

It turns out that *most neighbors of French, Spanish and Portuguese peak words are verbs*. Particularly, more than 97% of neighbors of Spanish and Portuguese peak words are verbs. *However, for Finnish, almost 90% of neighbors of peak words are nouns*.

Percentage (%)	French		Spanish		Portuguese		Finnish	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Words with types	97.39	0.88	96.96	1.73	97.74	0.75	93.72	4.45
Verb	87.26	25.85	97.24	2.59	99.23	0.94	3.86	14.64
Noun	9.34	20.08	2.07	2.15	0.77	0.94	89.67	26.50

Table 4.7: The mean and Standard Deviation (SD) percentage of verbs and nouns in the neighbors of peak words of network 'Related-To' in four inflecting languages.

This strengthens our belief that there may be a connection between the abnormal number of nodes with certain degree k and grammatical structures in these four languages. As we mentioned in Section 4.1, French, Spanish and Portuguese are classified as inflecting languages based on typological classification, here we explain two typical classes of inflection.

- *Conjugation*: inflection of verbs.
  Depending on the grammatical categories, the form of a verb changes. For example, 'slept' is an inflection of 'sleep'. Many italic languages, for example, Spanish, French and Portuguese, are rich in conjugations.

- *Declension*: inflection of nouns.
  In English, the declension is very simple, *e.g.*, the plural form 'men' is a declension of the singular form 'man'. In some other languages, declensions are more common and more diverse.

We suspect that high degrees of the peak words are related to the inflection in French, Spanish, Portuguese and Finnish. To validate our hypothesis, we make use of a new relation in ConceptNet: 'Form-Of'. The method and results are shown in the next section.

4.4.2. Merging of Inflected Words

To validate our hypothesis that the peaks in the degree distributions of these language networks are related to inflection, we inspect another relation in the dataset, 'Form-Of'. In ConceptNet, the relation 'Form-Of' connects two words A and B if A is an inflected form of B, or B is the root word of A [60].

Our idea is to merge every root word and its inflection (neighbors) in the 'Form-Of' network, and apply the merged words to the 'Related-To' network (see Fig. 4.3). Then we evaluate the effect of language inflection on the degree distribution. If the peak in the degree distribution disappears, it proves that the peak words are related to language inflection.

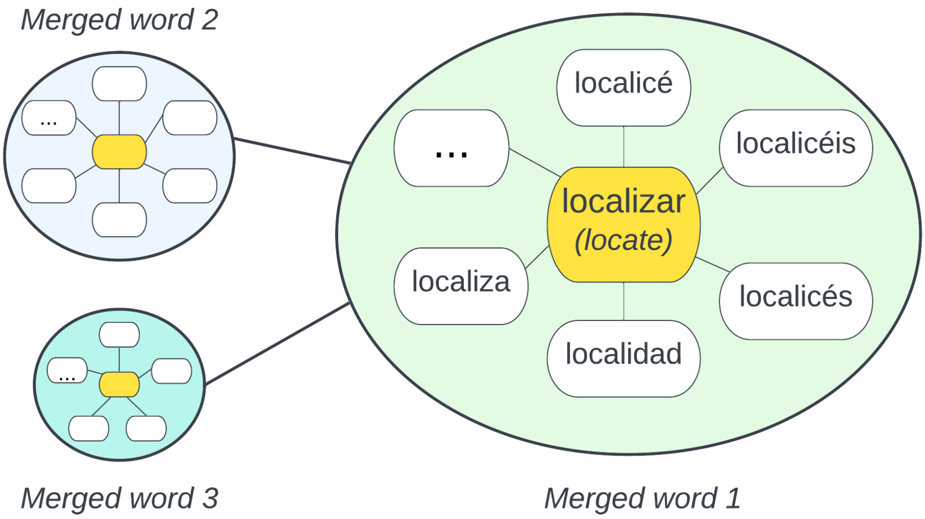


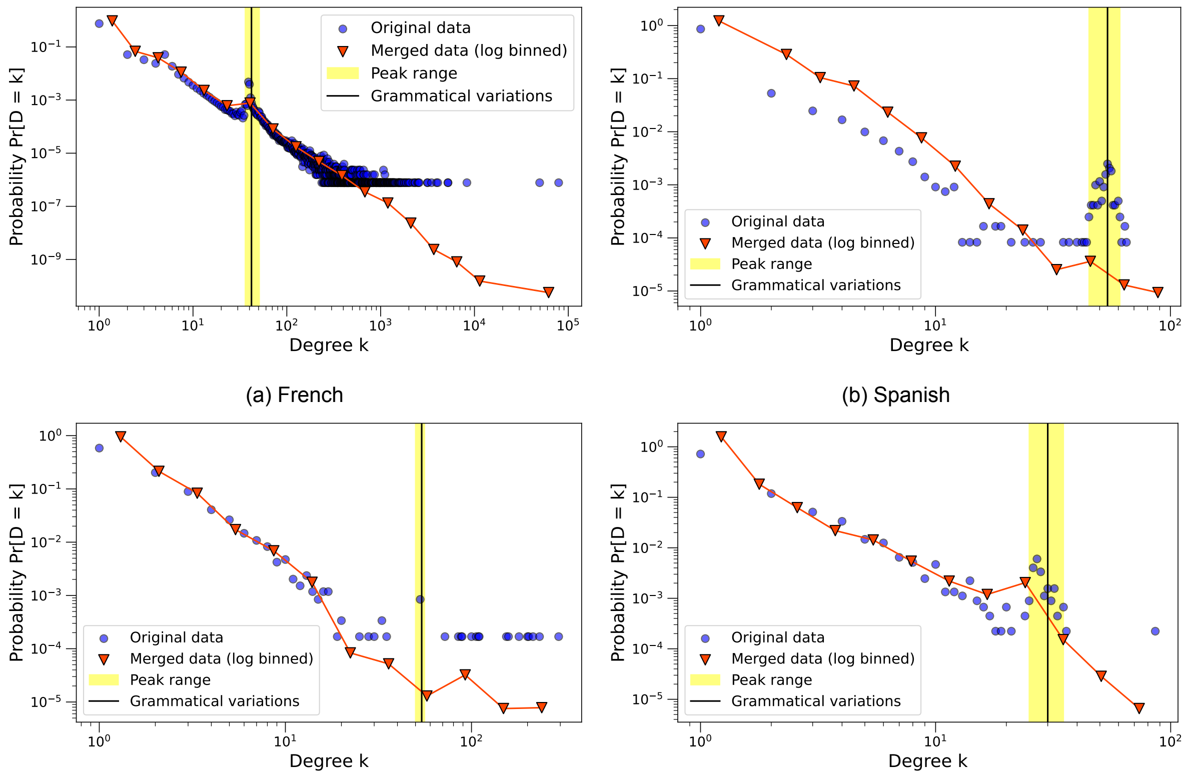
Figure 4.3: Illustration of words merging in the 'Related-To' network. After merging a root word and its neighbors, all words in a circle are seen as one single word.

First, we extract the network 'Form-Of' in the same way as for all other networks. Then we treat the merged group of words as a single word in the 'Related-To' network in the same language. Next we calculate the number of nodes with degree k in the new 'Related-To' network. Finally, we plot the degree distribution of French, Spanish, Portuguese and Finnish networks.

Fig. 4.4 illustrates the degree distribution of the original 'Related-To' network and after node merging. We highlight the anomalous peak in the degree distribution in yellow. These peaks indicate that there are more number of nodes with certain degree k than expected based on the power-law.

As shown in Fig. 4.4b, the peak completely disappears in Spanish 'Related-To'. This tells us that after merging the inflected words, there are no more number of nodes with degree k than expected. This validates our hypothesis that peak words and corresponding neighboring words are inflected forms, *i.e.*, conjugations.

We also observe the peak reduction in the degree distribution of Portuguese and Finnish ‘Related-To’. It seems that only some of the inflections are merged but not all. However, there seems no big change in the degree distribution of French ‘Related-To’. We believe this minor reduction of the peak is caused by the partial coverage in the ‘Form-Of’ network. That is the ‘Form-Of’ network does not contain all peak words and their inflected forms.



(a) French

(b) Spanish

(c) Portuguese

(d) Finnish

Figure 4.4: Degree distributions of the original networka ‘Related-To’ and after node merging in French, Spanish, Portuguese and Finnish. The logarithmically binned degree distributions of networks after node merging are shown in red. The peaks are highlighted in yellow. The vertical black lines indicate the number of grammatical variations in different languages, which are derived from corresponding grammatical rules (see Section 4.4.3).

To validate that network ‘Form-Of’ does not include all peak words and their inflected forms, we compute the number of peak words N_P and the number of common nodes N_C in the ‘Form-Of’ network and the peak words. Dividing N_P over N_C gives us the percentage of peak words covered by ‘Form-Of’. Similarly, we obtain the percentage of neighbors of peak words covered by ‘Form-Of’. The results are presented in Table 4.8.

We find that more than 97% of peak words and their neighbors are covered by ‘Form-Of’ in Spanish. This explains the disappearing of the peak in degree distribution. While there is only 17% of words matched by ‘Form-Of’ in the French ‘Related-To’ network. Hence, we see no big change in its degree distribution. As for Finnish and Portuguese, the percentage of matched words is moderate, which is around 50%. This is reflected in the minor reduction of the peak in the degree distribution.

Percentage	**French**	**Spanish**	**Portuguese**	**Finnish**
Percentage of peak words covered by ‘Form-Of’	33.72%	100%	60.00%	91.30%
Percentage of neighbors of peak words covered by ‘Form-Of’	17.38%	97.76%	55.47%	45.08%

Table 4.8: The percentage of matched words in peak words of network ‘Related-To’ in four languages.

4.4.3. Grammar

After investigating the percentages of verbs and nouns in the peak words, we are now convinced that the anomaly in degree distribution of ‘Related-To’ is closely related to language inflection. But does the location where the peaks occur correspond to the number of grammatical variations? To answer this question, we refer to the grammatical rules of these four languages and examine whether the degree range where the peaks appear matches those rules. We focus on the basic grammar, any irregular forms are not considered in this study.

To begin with, we introduce several important terms for better understanding of the grammatical rules.

- In grammar, a *pronoun* is a word that can substitute a noun. For example, ‘you’, ‘she’ and ‘they’ are pronouns in English.
- A *tense* is a grammatical time reference [70]. Tenses typically appear in specific forms of verbs (*e.g.*, conjugations). In English, typical tenses are past, present and future tense. A verb may change its form depending on the combination of the pronoun and tense.

We first look at the grammar for verbs. In Spanish, there are 6 pronouns and 9 simple verb tenses [71, 72]. The simple verb tenses are in the single word form. Each pronoun has its distinctive verb form (conjugation), and there are 54 (6 times 9) combinations of pronouns and tenses. As a result, the standard number of inflections of a Spanish verb is around 54. Table 4.9 provides an example of a Spanish verb in these tenses and pronouns.

Similar to Spanish, Portuguese has 6 pronouns and 9 tenses [73]. This results in 54 inflected forms of a Portuguese verb in general. As for French, there are 6 pronouns and 7 tenses (5 simple tenses and 2 mood tenses) [74]. There is another mood tense in French which has only a few verb variations, thus, we do not take this tense into account. Therefore, the number of inflections of a French verb is around 42.

In Finnish grammar, the form of a noun changes according to grammatical cases. There are in total 15 cases which are manifested in different endings at the nouns [75]. A noun has singular and plural forms. Thus there are approximately 30 inflected forms of a Finnish noun.

We now have an idea of the number of inflected forms m in these languages. The exact number of inflection forms of a word varies from case to case. Thus, we use our obtained m as a reference to the general number of grammatical variations.

Table 4.10 summarizes the number of grammatical variations m in basic French, Spanish, Portuguese and Finnish grammar. The minimum and maximum degree k_{min}

and k_{max} where the peak starts and ends are also listed. We can see that the numbers of grammatical variations m of these four languages land in or next to the range $[k_{min}, k_{max}]$. This further validates our hypothesis: the peak in the degree distribution of network 'Related-To' manifests language inflection.

Pronoun Tense	Yo (I)	Tú (You)	Él/Ella/Usted (He/She)	Nosotros (We)	Vosotros (You)	Ellos/Ellas /Ustedes (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 4.9: Conjugated forms of the Spanish verb 'amar' (to love) based on 6 pronouns and 9 tenses.

Language	Grammatical variations	k_{min}	k_{max}
French	42	36	51
Spanish	54	45	61
Portuguese	54	53	53
Finnish	30	25	35

Table 4.10: The number of grammatical variations for basic grammar in French, Spanish, Portuguese and Finnish. The minimum and maximum degree k_{min} and k_{max} where the peak starts and ends in the degree distributions of networks 'Related-To' are included for comparison.

Discussion Summarizing, *we observe grammatical features in the degree distributions of 'Related-To' networks*. Because of the special structure of French, Spanish, Portuguese and Finnish, words in these languages have many distinct inflections. There are more words with certain degree than expected. Consequently, we observe peaks in the degree distributions, which results in the deviation from a power-law. For French, Spanish and Portuguese, the inflected words are mostly conjugations. Whereas for Finnish, the inflected words are mostly declensions.

As defined in the typological classification of languages, French, Spanish and Portuguese are inflecting. However, although Finnish is typologically classified as agglutinating, it still has many declensions. This suggests that the two language categories (agglutinating and inflecting) are not mutually exclusive.

4.5. Degree Correlation Coefficient

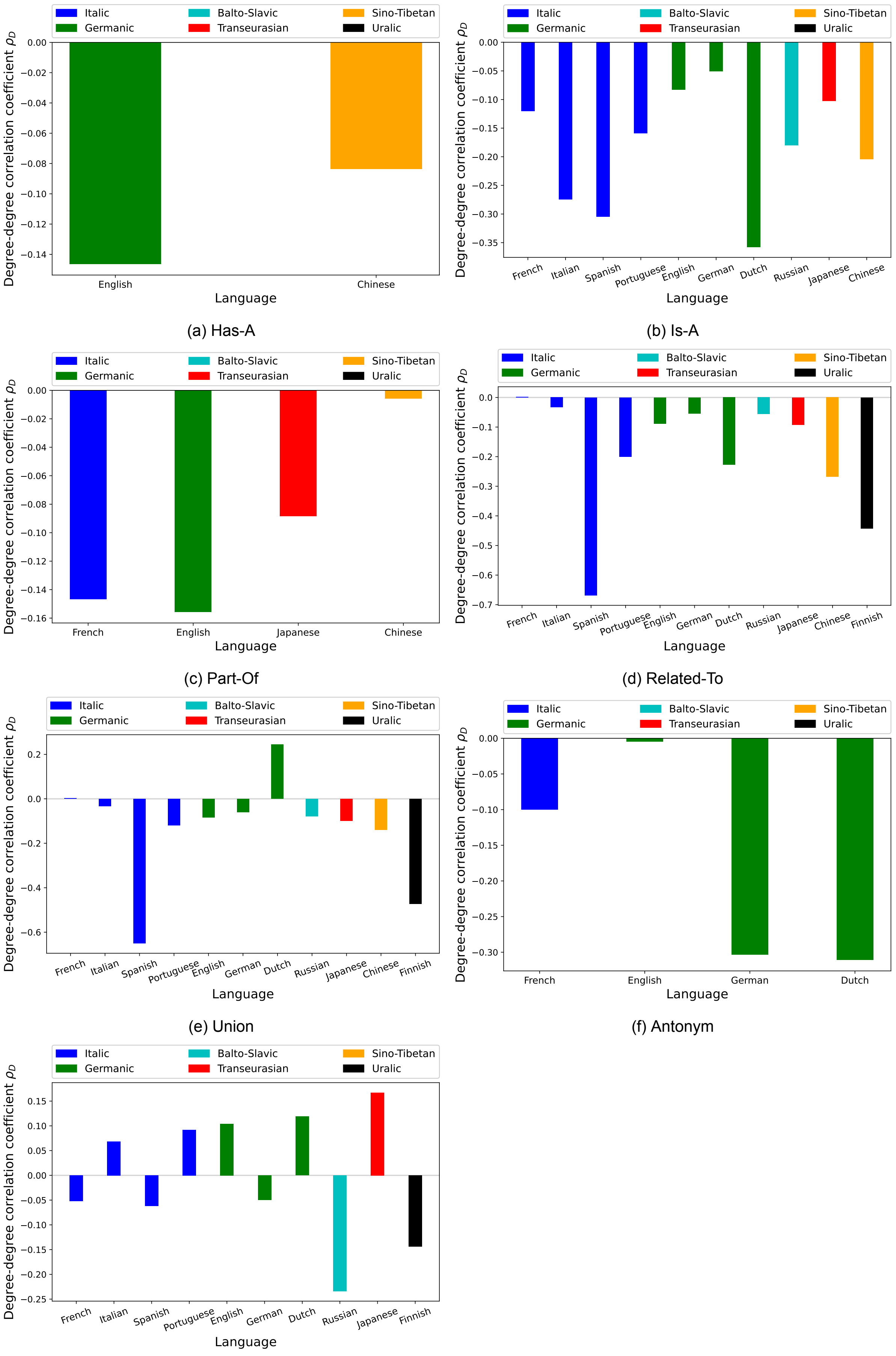
After exploring the degree distributions of semantic networks in different languages, we are now curious about the mixing patterns in them. Do networks in different languages have similar degree-degree mixing? Or are there differences between languages? Hence, we compute the degree correlation coefficient ρ_D of networks with size $N > 100$ using Eq. 2.7. The results for each network are presented in Appendix C.2.

To compare the degree assortativity of different languages and language families, we plot the results as bar charts in Fig. 4.5. Different colors represent different genetic language families.

Discussion The figure demonstrates that most networks have negative degree correlation coefficients. This coincides with our finding in English semantic networks in Section 3.7. Specifically, networks 'Has-A', 'Part-Of', 'Is-A' and 'Antonym' are disassortative regardless of the language. Additionally, in the 'Related-To' networks, only French has a slightly positive degree correlation coefficient.

However the signs of the degree correlation coefficients of network 'Synonym' vary from language to language. For example, the Russian 'Synonym' network has $\rho_D < -0.2$ while Japanese 'Synonym' has $\rho_D > 0.15$. Moreover, the degree assortativity differs even within the same language family, *e.g.*, the Italian and Spanish 'Synonym' networks.

In the 'Union' networks, only Dutch shows a clear positive degree correlation coefficient, but it is not immediately clear what is the cause. The French 'Union' network shows a slightly positive $\rho_D = 0.002$ because 'Related-to' makes up the majority of this network, and the French 'Related-To' network has positive degree correlation. As for 'Union' in the other languages, they present negative degree correlation.



(a) Has-A

(b) Is-A

(c) Part-Of

(d) Related-To

(e) Union

(f) Antonym

(g) Synonym

Figure 4.5: Degree correlation coefficient ρ_D of semantic networks in eleven languages (classified into six genetic language families). Only networks with size $N > 100$ are shown.

4.6. Clustering

Similarly, we study the clustering in these networks. Is high clustering a uniform property across semantic networks in different languages? To investigate this, we compute the average clustering coefficient and graph transitivity for networks with size $N > 100$. See numerical results in Appendix C.3 and C.4. Fig. 4.6 shows the average clustering coefficient of semantic networks in eleven languages. Languages that belong to the same genetic family share one color.

Discussion As shown in the figure, the average clustering coefficient varies for different networks in different languages. We observe small clustering coefficients in all ‘Antonym’ networks. Other than that, there are no obvious patterns in languages from the same family.

The graph transitivity of semantic networks in eleven languages are compared in Fig. 4.7. Most ‘Synonym’ networks have a larger graph transitivity $\check{c}_G > 0.10$ than other networks, which indicates that there are more triangles in ‘Synonym’. Besides, Dutch ‘Union’, Japanese ‘Related-To’ and Russian ‘Related-To’ and ‘Union’ also have larger graph transitivity $\check{c}_G > 0.10$ than others. The rest of the networks show a graph transitivity smaller than 0.1, particularly for ‘Antonym’. This reveals that in ‘Antonym’ networks, there are few connected triples. Likewise, we observe no clustering patterns in semantic networks from same language families.

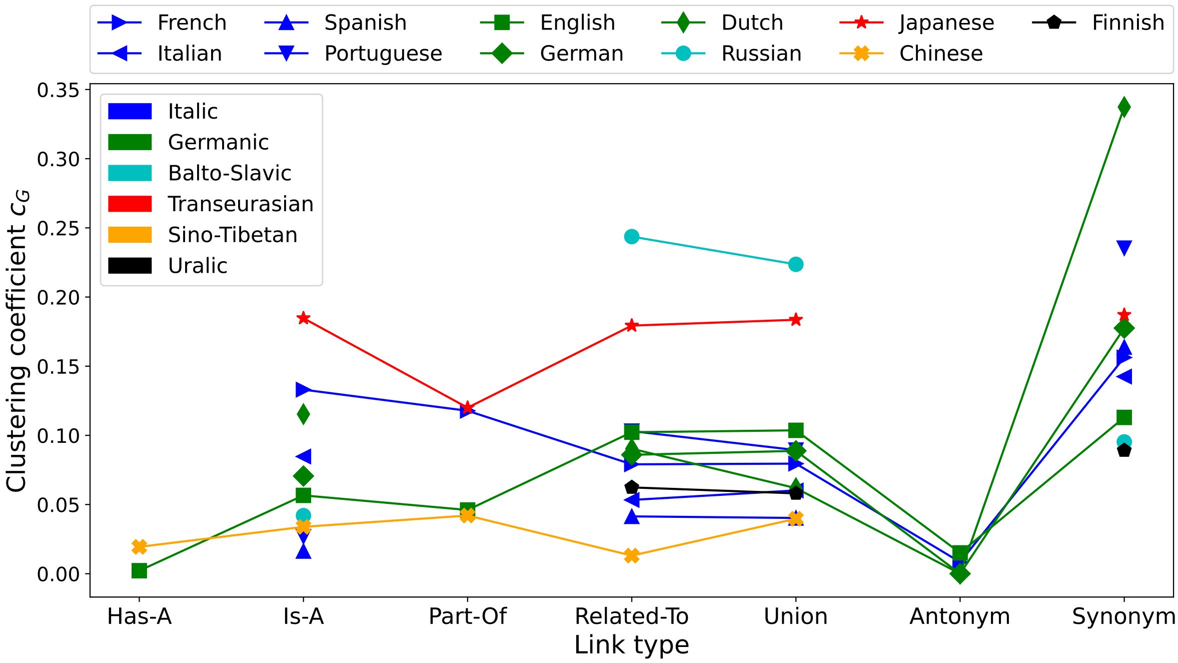


Figure 4.6: Average clustering coefficient c_G of seven semantic networks in eleven languages (classified into six genetic language families). Only networks with size $N > 100$ are shown.

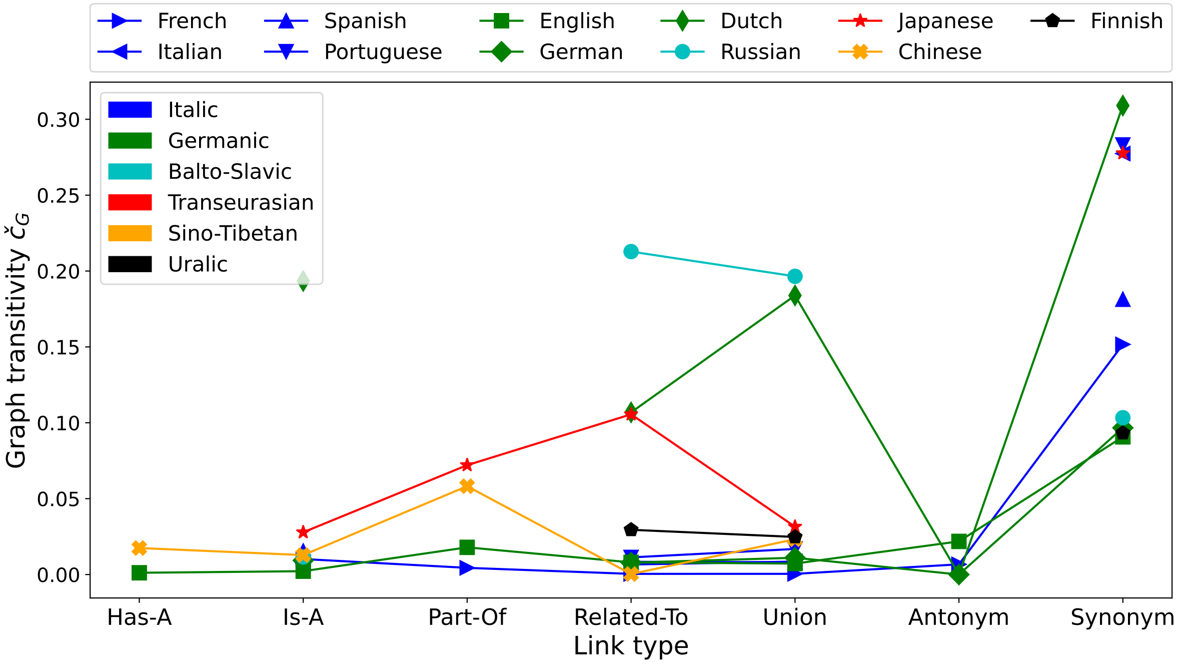


Figure 4.7: Graph transitivity $\check{c}_G$ of seven semantic networks in eleven languages (classified into six genetic language families). Only networks with size $N > 100$ are shown.

5

Similarity and Complementarity in Semantic Networks

In the previous chapters we study the basic topological properties of semantic networks. What are the organizing principles of semantic networks? This is the question that inspires this chapter. We already see from previous chapters that there are universal similarities between their network structures. But we also notice several structural differences. For example, some of the networks have higher clustering than rewired networks while others do not. Some networks are assortative but others are disassortative. We therefore conjecture that there are different mechanisms involved in the formation of these semantic networks.

A well-studied organizing principle is similarity. In networks where connections are driven by similarity, similar nodes (based on the number of common neighbors) are likely to be connected. A lot of NLP algorithms are developed based on social network analysis, where similarity is recognized as one of the leading mechanisms. In social networks, similar people are more likely to be connected.

Does this principle of similarity naturally make sense in our semantic networks? Do similar words co-occur in the same text or sentence? If we only use similar words in speech, it would sound rather boring. In reality, we combine different types of words, e.g., nouns and adjectives, verbs and nouns, and verbs in combination with adverbs. Different types of words complement each other and together form sentences. Additionally, when we compare things with contrasting nature, the words we use to describe them have contrasting meanings, too. All these cases we mention do not fall under the umbrella of similarity. In sentences, words have different meanings and types.

Hence, we presume that there is something else that drives the connections in semantic networks. Complementarity can be the other important organizing principle.

In this chapter we set to evaluate the organizing principles of semantic networks in a systematic way. First, we introduce the definition of similarity and complementarity, and the importance of them. Then, we present two measures that can quantify structural similarity and complementary of a network. Relying on these measures, we calculate the structural coefficients of our semantic networks and compare the results. At last, we conclude this chapter with important insights for improving NLP algorithms.

5.1. Similarity and Complementarity

Historically, the principle of **similarity** has been identified in various types of networks, including friendship, marriage and information exchange networks [24, 25]. Similarity is transitive. If node A is similar to B and B is similar to C, then A is similar to C. Thus, transitivity implies triangle closure (Fig. 5.1a) [25, 76, 77]. Due to the transitivity of similarity, there are lots of triangles in similarity-based networks. In a similarity-driven network, nodes that have many common neighbors are expected to be connected. For example, two persons who have a lot of common friends are very likely to be friends as well [78]. Many state-of-the-art algorithms in network science were developed based on this triangle closure principle [28].

However, in many other networks, nodes form connections not because they are similar but because they have complementary properties [79]. What is complementarity? Intuitively, two different objects complement each other by providing qualities or attributes that the other object lacks. The connection principle of **complementarity** is discovered in networks such as molecular interaction [80], interdisciplinary collaboration [79] and production networks [81]. For example, in company-level production networks, trading partners complement each other. Unlike similarity, complementarity is not transitive. If node A and B are complementary and B and C are complementary, it does not mean that A and C also complement each other. As a result, the triangle closure principle does not hold in complementarity-based networks. Instead, recently it was shown that complementarity-based connections lead to a large number of quadrangles in a network [82].

The emerging study of complementarity challenges the established methods of network science rooted in social networks. In a recent work on protein interactions [80], the authors show that current algorithms fail to accurately predict protein interactions due to the different organizing principles of protein interactions networks. Proteins interact when one of them is similar to the other's partners, not when they are similar to each other [80].

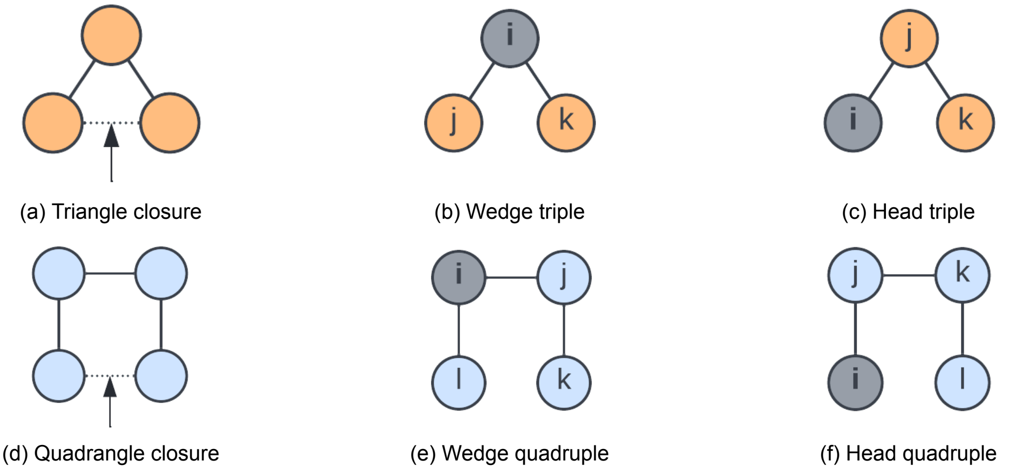
5.2. Structural Coefficients

From the previous section, we learn that similarity-based networks are rich in triangles because of the triangle closure principle. Our first observations are in the context of clustering coefficients in Section 3.8. The clustering coefficient is a classic measure of the density of triangles in a network. However, we cannot simply compare the number of triangles and quadrangles between two networks, because these networks have different sizes and degree distributions. We need to reliably calculate the statistics of triangles and quadrangles of a network to quantify similarity and complementarity. To this end, we rely on a recent work of complementarity [31]. Analogous to the clustering coefficient, we can use structural complementarity measures based on quadrangle closure rules (Fig. 5.1d).

Table 5.1 lists the procedures of how we compute the structural similarity and complementarity coefficients to quantify the density of triangles and quadrangles in a network G, respectively.

Procedure	Structural coefficients	Network G	
		Similarity ($\triangle$)	**Complementarity** ($\square$)
Step 1	Wedge triple/quadruple Head triple/quadruple	s_i^W, Eq. 5.1 s_i^H, Eq. 5.2	c_i^W, Eq. 5.5 c_i^H, Eq. 5.6
Step 2	Node-wise	s_i, Eq. 5.3	c_i, Eq. 5.7
Step 3	Network-wise	$s(G) = \frac{1}{N}\sum_{i=1}^{N} s_i$, Eq. 5.4	$c(G) = \frac{1}{N}\sum_{i=1}^{N} c_i$, Eq. 5.8
Step 4	Calibrated Network-wise	$\mathcal{C}(s)_G = \frac{1}{R}\sum_{i=1}^{R} \log \frac{s(G)}{s(G_i)}$	$\mathcal{C}(c)_G = \frac{1}{R}\sum_{i=1}^{R} \log \frac{c(G)}{c(G_i)}$

Table 5.1: The procedure of calculating the structural similarity coefficient and complementarity coefficient of a network G. The calibrated structural coefficients in step 4 are obtained by taking the average log ratio of network-wise coefficient over the coefficient of a sampled network G_i, see Eq. 5.9 in Section 5.3.1 and Appendix E.



(a) Triangle closure (b) Wedge triple (c) Head triple

(d) Quadrangle closure (e) Wedge quadruple (f) Head quadruple

Figure 5.1: Quadrangle and quadruples in comparison with triangle and triples. Wedge and head triples (or quadruples) are different at where node i is centered. Node i in a wedge triple (b) is centered in the middle, while i in a head triple (c) is centered at the beginning. Similarly, node i in a wedge quadruple (e) is centered at the second location, while i is at the beginning of a head quadruple (f).

5.2.1. Structural Similarity Coefficient

The structural similarity generalizes the local clustering and closure coefficients. The local clustering coefficient s_i^W of a node i is the same as $c_G(i)$ in Eq. 2.10. It is defined as the fraction of triples centered at i which can be closed to form a triangle [31]

$$s_i^W = \frac{2T_i}{t_i^W} = \frac{\sum_{j,k} a_{ij}a_{ik}a_{jk}}{d_i\,(d_i - 1)}, \tag{5.1}$$

where T_i is the number of triangles including i and t_i^W is the number of wedge triples (Fig. 5.1b), or 2-paths with node i in the middle, *e.g.*, (j, i, k). The definition of the local closure coefficient [83] is given as follows

$$s_i^H = \frac{2T_i}{t_i^H} = \frac{\sum_{j,k} a_{ij} a_{ik} a_{jk}}{\sum_j a_{ij} \left(d_j - 1\right)}, \tag{5.2}$$

where t_i^H is the number of head triples (Fig. 5.1c), *i.e.*, 2-paths starting from node i, such as (i, j, k). Both s_i^W and s_i^H are bounded in the range $[0, 1]$, but they capture different parts of the spectrum of similarity-driven structures [31].

Combining the weighted average of these two coefficients results in a more comprehensive measure of local structure, the *structural similarity coefficient* [31], which captures the full spectrum of structural similarity. It is defined as

$$s_i = \frac{4T_i}{t_i^W + t_i^H} = \frac{t_i^W s_i^W + t_i^H s_i^H}{t_i^W + t_i^H}. \tag{5.3}$$

The coefficient $s_i = 1$ only if node i is in a fully connected network.

The structural similarity coefficient of a whole network G is then the average over all nodes

$$s(G) = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{5.4}$$

5.2.2. Structural Complementarity Coefficient

Analogously, the local quadruples clustering coefficient at node i is defined as the fraction of closed quadruples with i at the second position [31]

$$c_i^W = \frac{2Q_i}{q_i^W} = \frac{\sum_{j \neq i} a_{ij} \sum_{k \neq i,j} a_{ik} \left(1 - a_{jk}\right) \sum_{l \neq i,j,k} a_{kl} a_{jl} \left(1 - a_{il}\right)}{\sum_j a_{ij} \left[\left(d_i - 1\right)\left(d_j - 1\right) - n_{ij}\right]}, \tag{5.5}$$

where Q_i represents the number of quadrangles contain that node i and q_i^W is the number of wedge quadruples (Fig. 5.1e), or 3-paths with i at the second node, *e.g.*, (l, i, j, k). Similarly, the local quadruples closure coefficient of a node i calculates the percentage of closed quadruples beginning at i

$$c_i^H = \frac{2Q_i}{q_i^H} = \frac{\sum_{j \neq i} a_{ij} \sum_{k \neq i,j} a_{ik} \left(1 - a_{jk}\right) \sum_{l \neq i,j,k} a_{kl} a_{jl} \left(1 - a_{il}\right)}{\sum_{j \neq i} a_{ij} \sum_{k \neq i,j} a_{jk} \left(d_k - 1 - a_{ik}\right)}, \tag{5.6}$$

where q_i^H is the number of head quadruples originating from node i (Fig. 5.1f).

Finally, the *structural complementarity coefficient* is constructed as the weighted average of the local quadruples clustering and closure coefficients [31]

$$c_i = \frac{4Q_i}{q_i^W + q_i^H} = \frac{q_i^W c_i^W + q_i^H c_i^H}{q_i^W + q_i^H}. \tag{5.7}$$

The structural complementarity coefficient $c_i \in [0, 1]$, which is proven to be a more general measure than using only c_i^W or c_i^H [31]. The maximum $c_i = 1$ happens only if node i belongs to a fully connected bipartite graph. In a bipartite graph, nodes are divided into two groups, and connections are only formed between groups but not within the same group.

The structural complementarity coefficient of a whole network G is then the average of all nodes:

$$c(G) = \frac{1}{N}\sum_{i=1}^{N} c_i. \tag{5.8}$$

5.3. Structural Similarity and Complementarity Coefficients in Semantic Networks

Using the structural similarity and complementarity coefficients, we measure and compare the density of triangles and quadrangles in a real network. Therefore, we can determine the relative roles of similarity and complementarity in a network. In this section, we calculate and compare the average structural coefficients of our semantic networks with the help of the algorithm provided in [31]. To compare the structural coefficients of different semantic networks, the values need to be normalized based on a configuration model to correct for the effects purely induced by the degree sequences [31]. The configuration model and more details of the calibration process are explained in Appendix E.

5.3.1. Calibration of Structural Coefficients

First of all, we calculate one structural coefficient (similarity or complementarity) of a given network G. We denote this coefficient as $x(G)$. Second, we sample R randomized copies G_i's of the given network from the Configuration Model (CM). Then, we calculate the structural coefficient $x(G_i)$ for each sampled network. At last, we take the average log-ratio of $x(G)$ and $x(G_i)$'s. As a result, the calibrated coefficient $\mathcal{C}(x)_G$ according to R samples from CM is obtained as follows [31]

$$\mathcal{C}(x)_G = \frac{1}{R}\sum_{i=1}^{R} \log \frac{x(G)}{x\left(G_i\right)}. \tag{5.9}$$

The calibrated structural coefficient can be less, equal or larger than zero. Consider the calibrated structural similarity coefficient $\mathcal{C}(s)_G$ for example:

- $\mathcal{C}(s)_G < 0$, the structural similarity coefficient $s(G)$ is smaller than $s(G_i)$ of random networks.
- $\mathcal{C}(s)_G = 0$, the structural similarity coefficient is comparable to the ones in random networks.
- $\mathcal{C}(s)_G > 0$, the structural similarity coefficient is larger than in random networks.

Fig. 5.2 shows the scatter plot of the calibrated average structural coefficients of 50 semantic networks in different languages. The grey lines at $x = 0$ and $y = 0$ indicate the expected coefficients based on the Configuration Model.

Discussion As illustrated in Fig. 5.2, we see clusters of colors but not really based on shapes. This suggests that the type of relation matters more for the organizing principles of a semantic network, rather than the language.

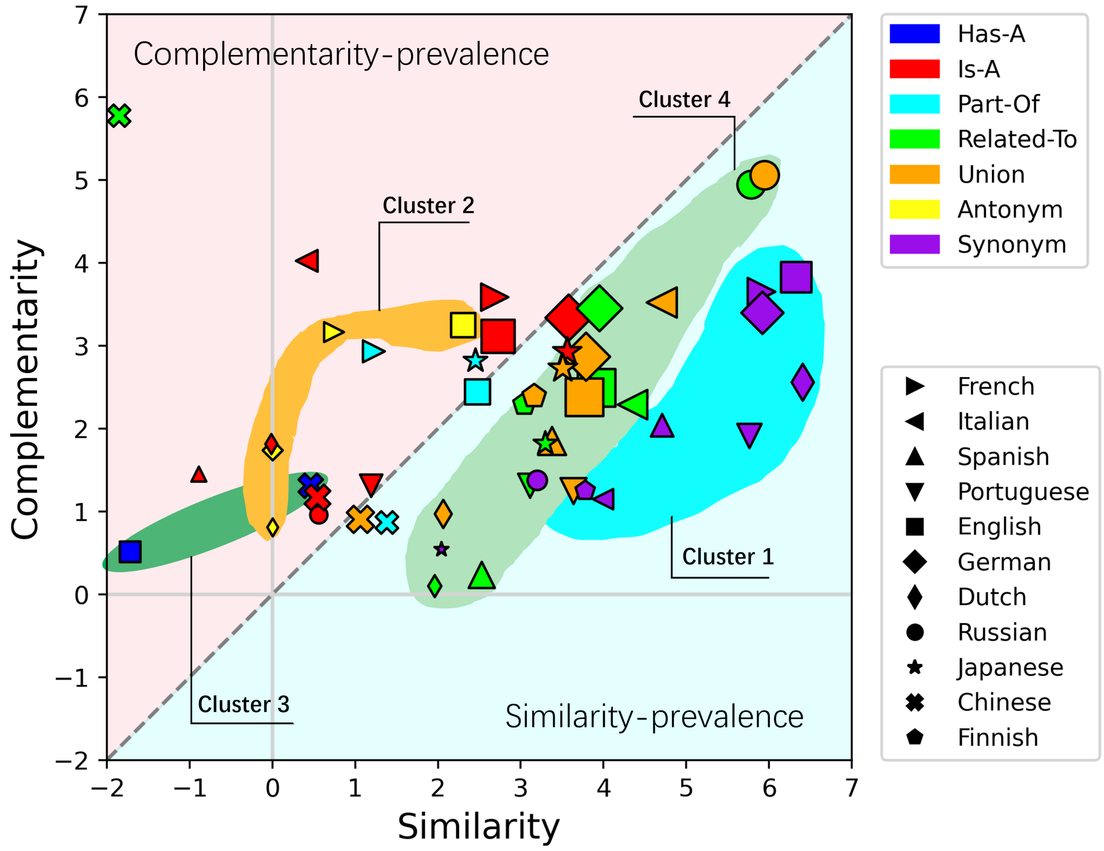


Figure 5.2: Calibrated average structural coefficients of 50 semantic networks in different languages. Languages that belong to the same family are marked with similar shapes. Triangles represent Italic, quadrilaterals represent Germanic, circles represent Balto-Slavic, star represents Transeurasian, cross represents Sino-Tibetan and pentagon represents Uralic. The marker size is proportional to $(\log(N))^{2.1}$, where N is the number of nodes in a network. The grey lines at $x = 0$ and $y = 0$ indicate the expected coefficients based on the configuration model (see Appendix E). The dashed line at $y = x$ indicates that the structural similarity and complementarity coefficients are equal. Networks in the upper left area (shaded in red) are more complementarity-based, while networks in the lower right area (shaded in blue) are more similarity-based. We highlight four clusters of networks using different colors.

If we inspect networks in specific languages, we observe symbols of the same shape distributed all over the plot. The fact that networks from one language do not cluster together implies that our networks are properly constructed, networks with different relation types show different connection principles.

There are mainly three categories of networks: predominantly complementarity-based, predominantly similarity-based, or both. Most semantic networks exhibit both patterns of similarity and complementarity. Our results show that the prevalence of similarity and complementarity in semantic networks is mostly related to semantic relation type. We list the observed patterns.

- Cluster 1 (cyan): *‘Synonym’ networks show stronger similarity than complementarity.* A cluster of ‘Synonym’ networks indicates that words tend to connect to words with similar meanings, which coincides with the definition of the ‘Synonym’ relation in ConceptNet. Consequently, there are lots of closed triples in these networks.

- Cluster 2 (orange): *‘Antonym’ networks are more complementarity-based*, that is, two words that share a lot of neighboring words are not certainly connected. As a result, there exist many quadrangles in these networks. This can be explained by two words that have opposite meanings of another word are not necessarily opposite to each other. For example, in the English ‘Antonym’ network, the word ‘small’ is the opposite of ‘big’ as well as ‘great’, however, ‘big’ and ‘great’ do not have opposite meanings. But since the word ‘little’ is also connected to ‘big’ and ‘great’, the four words form a quadrangle.

- Cluster 3 (dark green): *‘Has-A’ networks show more structural complementarity than similarity.* Intuitively, words in ‘Has-A’ complement one another. For instance, ‘a *house* has a *roof*’ is a complementary relation, while these two objects are not similar to each other..

- Cluster 4 (light green): *Most ‘Related-To’ and ‘Union’ networks show more similarity, except for Chinese.* As defined in the ‘Related-To’ relation, words are connected if there is a positive relation between them, therefore, it is easy to form triangles.

  Though one exception is that Chinese ‘Related-To’ (green cross) shows the strongest complementarity among all networks and decreased similarity (relative to the configuration model). One possible explanation is that Chinese has plenty of measure words that are connected to all kinds of words (nouns). Measure words, also known as numeral classifiers, are used in combination of numerals to describe the quantity of things [84, 85]. For example, in English we usually say ‘one apple’, but in Chinese a measure word must be added between the number ‘one’ and the noun ‘apple’ as a unit of measurement. This grammatical phenomenon is comparable to when we say ‘one box of apples’ in English, but in English these measure words are rare. Depending on the situation, the measure word of the same noun can be different. In the Chinese ‘Related-To’ network, there are many measure words connecting to different nouns. But these nouns may have no connection between each other at all. Hence, the structural similarity is lower than expected from the configuration model. However, since there are many choices for the measure word of a noun and one measure word can be used with multiple nouns, it leads to numerous quadrangles. Therefore, Chinese ‘Related-To’ shows the highest structural complementarity coefficient.

- *Most large semantic networks present stronger complementarity and similarity.*

Conclusion To summarize, the connections in semantic networks are driven by similarity and/or complementarity, mainly depending on the semantic relation type. Networks from different languages may present very different complementarity-based structures due to grammatical features.

Since Natural Language Processing (NLP) methods use tools that are based on similarity, they may work very well for similarity-based semantic networks. But they are not expected to work as well for complementarity-based networks. Therefore, new NLP methods need to be developed accordingly.

6 Conclusion and Future Work

In this chapter, we review the main objective of this thesis and summarize the study that has been carried out. Based on our results of structural properties of semantic networks, we draw conclusions and propose some suggestions for future work.

6.1. Conclusions

The primary objective of this thesis is to study the topological properties of semantic networks. We studied semantic networks with 7 distinct semantic relations from 11 different languages.

Overall, we observed universal characteristics in the basic structure of semantic networks. In chapter 3, we focused on the study of seven English semantic networks: ‘Has-A’, ‘Part-Of’, ‘Is-A’, ‘Related-To’, ‘Union’, ‘Antonym’ and ‘Synonym’. We found that these semantic networks can be characterized with high sparsity and a power-law degree distribution. We also found that most semantic networks are scale-free. We observed two patterns of degree mixing in these networks. Some networks are assortative and others are disassortative. In addition, we found that most networks have higher clustering coefficients than degree-preserving rewired networks.

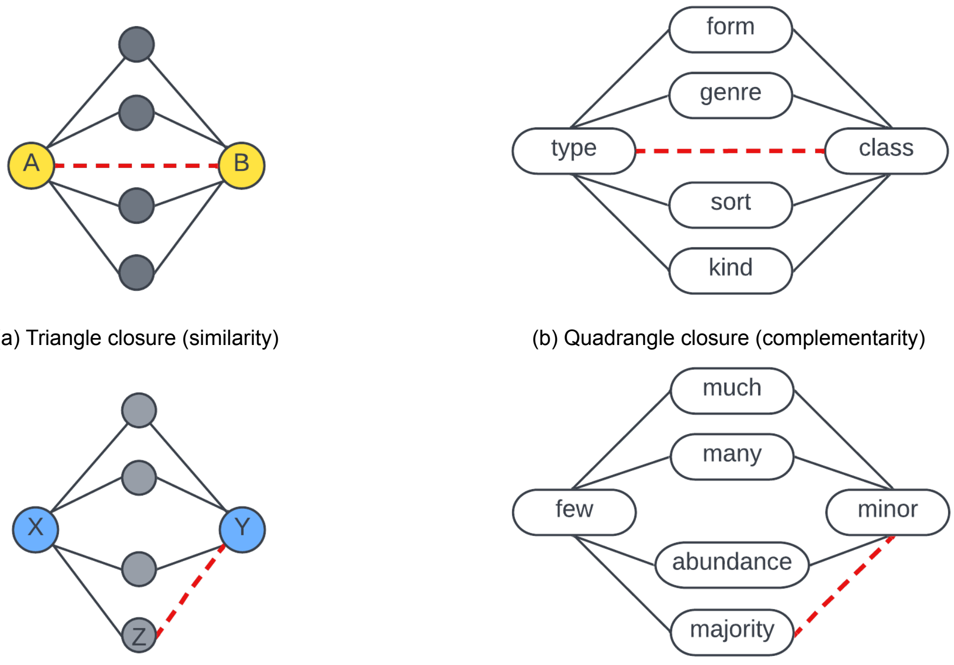
On the other hand, we also found different properties in semantic networks from different languages. In Chapter 4, we considered semantic networks from 11 languages. They are English, French, Italian, German, Spanish, Russian, Portuguese, Dutch, Japanese, Finnish and Chinese. We divided them into different language families according to two classifications: typological and genetic. Interestingly, we discovered non-trivial structural patterns in networks from languages that have many grammatical inflections, *i.e.*, French, Spanish, Portuguese and Finnish. Because of the natural structure of grammar in these languages, words have a large number of conjugations or declensions. A large number of connections due to inflections results in peaks in the degree distributions. Moreover, we found that not only inflecting languages have many inflected forms of words but also one agglutinating language, which is Finnish.

All the aforementioned structural patterns in semantic networks encouraged us to investigate the organizing principles of these networks. We introduced the two organizing principles similarity and complementarity in Chapter 5. By computing the structural similarity and complementarity coefficients of 50 semantic networks from different languages, we observed both similarity and complementarity in the connection principles of semantic networks. But the proportions of similarity and complementarity in networks are different depending on the semantic relation type. For example, 'Synonym' networks show stronger similarity, while connections in 'Antonym' are more driven by complementarity. Additionally, the Chinese 'Related-To' network stood out with the highest structural complementarity coefficient from the rest of the networks. We were able to partially relate this strong complementarity to a unique grammatical phenomenon in Chinese: measure words.

The results we presented in Chapter 5 are important for Natural Language Processing (NLP), because NLP algorithms mostly rely on the similarity principle and neglect the principle of complementarity. Existing NLP algorithms may work well for networks that are similarity-based, but different methods are required for processing complementarity-based semantic networks.

6.2. Outlook

The motivation of our study was the desire to improve upon existing NLP technologies. Though we did not give an exact solution, we are certain that our results (especially in Chapter 5) serve as evidence that we inform better NLP methods. Here we give an example of where the difference of similarity and complementarity manifests in a fundamental task, that is, *link prediction*.



(a) Triangle closure (similarity)

(b) Quadrangle closure (complementarity)

(c) Quadrangle closure (complementarity)

(d) Quadrangle closure (complementarity)

Figure 6.1: Comparison of similarity and complementarity principles in networks. (a) Lots of common neighbors of A and B imply similarity between A and B, therefore they are connected. (b) An example of triangle closure in 'Synonym' network. (c) In complementarity-based networks, if node X and Y share many common neighbors, the additional neighbor Z of node X implies the link between Z and Y. (d) An example of quadrangle closure in 'Antonym' network.

In the view of traditional link prediction (see Fig. 6.1a), two nodes A and B are considered to be similar if they have a lot of common neighbors. Therefore, nodes A and B must be connected. The basic rational is that people who have many common friends will most likely also establish connections. In the ‘Synonym’ network (Fig. 6.1b), the words ‘type’ and ‘class’ share lots of neighbors that have similar meanings, such as ‘kind’, ‘form’ and ‘genre’. Therefore, ‘type’ and ‘class’ also have similar meanings.

However, in complementarity-driven networks, the principle of similarity does not work for predicting links. Two nodes that share a lot of common neighbors maybe similar, but they do not necessarily complement each other. Instead, the connection principle is, if a node X has an additional connection to a node Z that is not shared with node Y, then node Z and Y might be connected as well (See Fig. 6.1c). For example, in our ‘Antonym’ network (Fig. 6.1d), the words ‘few’ and ‘minor’ have lots of neighbors that have the opposite meaning as them. The word ‘few’ is additionally connected to ‘majority’. Hence, we can conclude that ‘majority’ and ‘minor’ are also connected by the antonym relation.

The principle of complementarity is not only identified in semantic networks but also in other network classes such as interdisciplinary collaboration [79], biological [80] and company-level production networks [81]. And traditional algorithms such as link prediction fail to yield satisfying results [80]. Therefore, it is important to develop algorithms that take complementarity into account.

We identify the following interesting directions for future research on semantic networks, NLP and related fields.

- It is recommended to investigate semantic networks extracted from other databases to compare to the topological properties we found in ConceptNet.
- To obtain a better overview of differences and similarities between semantic networks from different languages, it is also suggested to further explore the topological properties of semantic networks in the same language family and between different language families.
- Our results imply that some of the existing similarity-based methods need to be revised. It is recommended to study basic NLP methods, such as link prediction and sentiment analysis, to gain more insights of how to reformulate these methods.

Bibliography

[1] Erik Cambria and Bebo White. “Jumping NLP curves: A review of natural language processing research”. In: *IEEE Computational intelligence magazine* 9.2 (2014), pp. 48–57.

[2] John F Sowa. “Semantic networks”. In: *Encyclopedia of Cognitive Science* (2012).

[3] Hernane Borges de Barros Pereira et al. “Systematic review of the “semantic network” definitions”. In: *Expert Systems with Applications* (2022), p. 118455.

[4] M Ross Quillian. “Word concepts: A theory and simulation of some basic semantic capabilities”. In: *Behavioral science* 12.5 (1967), pp. 410–430.

[5] M Ross Quillian. “The teachable language comprehender: A simulation program and theory of language”. In: *Communications of the ACM* 12.8 (1969), pp. 459–476.

[6] John F Sowa. *Principles of semantic networks: Explorations in the representation of knowledge*. Morgan Kaufmann, 2014.

[7] Stephen Peters and Howard E Shrobe. “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).* IEEE. 2003, pp. 323–329.

[8] Abdel-Badeeh M Salem and Marco Alfonse. “Ontology versus semantic networks for medical knowledge representation”. In: *Recent Advances In Computer Engineering* (2008), pp. 769–774.

[9] Amit Singhal. *Introducing the Knowledge Graph: things, not strings*. May 2012. URL: `https://blog.google/products/search/introducing-knowledge-graph-things-not/`.

[10] Roel Popping. “Knowledge graphs and network text analysis”. In: *Social Science Information* 42.1 (2003), pp. 91–106.

[11] Dieter Fensel et al. “Introduction: what is a knowledge graph?” In: *Knowledge Graphs*. Springer, 2020, pp. 1–10.

[12] Veton Kepuska and Gamal Bohouta. “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)*. IEEE. 2018, pp. 99–103.

[13] Rob High. “The era of cognitive systems: An inside look at IBM Watson and how it works”. In: *IBM Corporation, Redbooks* 1 (2012), p. 16.

[14] Jakub Piskorski and Roman Yangarber. “Information extraction: Past, present and future”. In: *Multi-source, multilingual information extraction and summarization*. Springer, 2013, pp. 23–49.

[15] Feng Shi et al. “A data-driven text mining and semantic network analysis for design information retrieval”. In: *Journal of Mechanical Design* 139.11 (2017).

[16] Philip Resnik. “Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language”. In: *Journal of artificial intelligence research* 11 (1999), pp. 95–130.

[17] Ana Harris. *Human languages vs. programming languages*. Nov. 2018. URL: `https://medium.com/@anaharris/human-languages-vs-programming-languages-c89410f13252`.

[18] Diksha Khurana et al. “Natural language processing: State of the art, current trends and challenges”. In: *Multimedia Tools and Applications* (2022), pp. 1–32.

[19] J Martijn Nobel et al. “Natural language processing in Dutch free text radiology reports: challenges in a small language area staging pulmonary oncology”. In: *Journal of digital imaging* 33.4 (2020), pp. 1002–1008.

[20] Hejab M Alfawareh and Shaidah Jusoh. “Resolving ambiguous entity through context knowledge and fuzzy approach”. In: *International journal on computer science and engineering (IJCSE)* 3.1 (2011), pp. 410–422.

[21] Shaidah Jusoh. “A study on NLP applications and ambiguity problems.” In: *Journal of Theoretical & Applied Information Technology* 96.6 (2018).

[22] Kareem Darwish et al. “A panoramic survey of natural language processing in the Arab world”. In: *Communications of the ACM* 64.4 (2021), pp. 72–81.

[23] Daniel Hershcovich et al. “Challenges and strategies in cross-cultural NLP”. In: *arXiv preprint arXiv:2203.10020* (2022).

[24] Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a feather: Homophily in social networks”. In: *Annual review of sociology* (2001), pp. 415–444.

[25] Gueorgi Kossinets and Duncan J Watts. “Origins of homophily in an evolving social network”. In: *American journal of sociology* 115.2 (2009), pp. 405–450.

[26] David R Schaefer et al. “Fundamental principles of network formation among preschool children”. In: *Social networks* 32.1 (2010), pp. 61–71.

[27] Tom AB Snijders. “Statistical models for social networks”. In: *Annual review of sociology* 37 (2011), pp. 131–153.

[28] Mohammad Al Hasan and Mohammed J Zaki. “A survey of link prediction in social networks”. In: *Social network data analytics*. Springer, 2011, pp. 243–275.

[29] Fataneh Dabaghi Zarandi and Marjan Kuchaki Rafsanjani. “Community detection in complex networks using structural similarity”. In: *Physica A: Statistical Mechanics and its Applications* 503 (2018), pp. 882–891.

[30] Nicholas Evans and Stephen C Levinson. “The myth of language universals: Language diversity and its importance for cognitive science”. In: *Behavioral and brain sciences* 32.5 (2009), pp. 429–448.

[31] Szymon Talaga and Andrzej Nowak. *Structural complementarity and similarity: linking relational principles to network motifs*. 2022. DOI: `10.48550/ARXIV.2201.03664`.

[32] P Zhang and G Chartrand. *Introduction to graph theory*. Tata McGraw-Hill, 2006.

[33] Piet Van Mieghem. *Graph spectra for complex networks*. Cambridge University Press, 2010.

[34] Albert-László Barabási. “Network science”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1987 (2013), p. 20120375.

[35] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *science* 286.5439 (1999), pp. 509–512.

[36] Mark EJ Newman. “Mixing patterns in networks”. In: *Physical review E* 67.2 (2003), p. 026126.

[37] Marián Boguñá and Romualdo Pastor-Satorras. “Class of correlated random networks with hidden variables”. In: *Physical Review E* 68.3 (2003), p. 036112.

[38] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), pp. 440–442.

[39] Piet Van Mieghem. *Performance analysis of complex networks and systems*. Cambridge University Press, 2014.

[40] Stanley Wasserman, Katherine Faust, et al. “Social network analysis: Methods and applications”. In: (1994).

[41] Mark EJ Newman. “The structure of scientific collaboration networks”. In: *Proceedings of the national academy of sciences* 98.2 (2001), pp. 404–409.

[42] Krishnaiyan Thulasiraman and Madisetti NS Swamy. *Graphs: theory and algorithms*. John Wiley & Sons, 2011.

[43] Paul Erdős, Alfréd Rényi, et al. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1 (1960), pp. 17–60.

[44] Mark Newman. *Networks*. Oxford university press, 2018.

[45] DL Nelson. “The University of South Florida word association norms”. In: *http://w3. usf. edu/FreeAssociation* (1999).

[46] Ramon Ferrer-i-Cancho and Richard V Solé. “The small world of human language”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1482 (2001), pp. 2261–2265.

[47] BNC Consortium et al. “British national corpus”. In: *Oxford Text Archive Core Collection* (2007).

[48] Adilson E Motter et al. “Topology of the conceptual network of language”. In: *Physical Review E* 65.6 (2002), p. 065102.

[49] Grady Ward. *Moby thesaurus list*. Quality Classics, 2015.

[50] Mariano Sigman and Guillermo A Cecchi. “Global organization of the Wordnet lexicon”. In: *Proceedings of the National Academy of Sciences* 99.3 (2002), pp. 1742–1747.

[51] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[52] Christiane Fellbaum. *WordNet: An electronic lexical database and some of its applications*. 1998.

[53] Mark Steyvers and Joshua B Tenenbaum. “The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth”. In: *Cognitive science* 29.1 (2005), pp. 41–78.

[54] Peter Mark Roget. *Roget's Thesaurus of English Words and Phrases*. TY Crowell Company, 1911.

[55] Steven H Strogatz. “Exploring complex networks”. In: *nature* 410.6825 (2001), pp. 268–276.

[56] Ramon Ferrer-i-Cancho, Ricard V Solé, and Reinhard Köhler. “Patterns in syntactic dependency networks”. In: *Physical Review E* 69.5 (2004), p. 051915.

[57] Princeton University. *About WordNet*. 2010. URL: `https://wordnet.princeton.edu/`.

[58] Jens Lehmann et al. “Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia”. In: *Semantic web* 6.2 (2015), pp. 167–195.

[59] Robyn Speer, Joshua Chin, and Catherine Havasi. “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 4444–4451.

[60] Robyn Speer. *Conceptnet5 Wiki - Commonsense: Relations*. June 2019. URL: `https://github.com/commonsense/conceptnet5/wiki/Relations`.

[61] Javier Borge-Holthoefer and Alex Arenas. “Semantic networks: Structure and dynamics”. In: *Entropy* 12.5 (2010), pp. 1264–1302.

[62] Lada A Adamic and Bernardo A Huberman. “Power-law distribution of the world wide web”. In: *science* 287.5461 (2000), pp. 2115–2115.

[63] Hawoong Jeong et al. “The large-scale organization of metabolic networks”. In: *Nature* 407.6804 (2000), pp. 651–654.

[64] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. “On power-law relationships of the internet topology”. In: *ACM SIGCOMM computer communication review* 29.4 (1999), pp. 251–262.

[65] Mark EJ Newman. “Clustering and preferential attachment in growing networks”. In: *Physical review E* 64.2 (2001), p. 025102.

[66] Ernesto Estrada. “When local and global clustering of networks diverge”. In: *Linear Algebra and its Applications* 488 (2016), pp. 249–263.

[67] John Lyons. *Language classification*. URL: `https://www.britannica.com/science/linguistics/Other-relationships`.

[68] David M Eberhard, Gary F Simons, and Charles D Fennig. *Ethnologue: Languages of the World*. Dallas, Texas, 2022. URL: `http://www.ethnologue.com`.

[69] Martin Haspelmath. “The morph as a minimal linguistic form”. In: *Morphology* 30.2 (2020), pp. 117–134.

[70] Bernard Comrie. *Aspect: An introduction to the study of verbal aspect and related problems*. Vol. 2. Cambridge University Press, 1976.

[71] Christopher Kendris and Theodore Kendris. *501 Spanish verbs*. Barrons Educational Series, 2020.

[72] FRANCISCO J. VARE. *Your All-in-one Guide to the 18 Spanish Tenses and Moods*. July 2022. URL: `https://www.fluentu.com/blog/spanish/spanish-tenses/#toc_15`.

[73] John J Nitti and Michael J Ferreira. *501 portuguese verbs*. Simon and Schuster, 2015.

[74] Laura K Lawless. *The Everything French Verb Book: A Handy Reference for Mastering Verb Conjugation*. Simon and Schuster, 2005.

[75] Fred Karlsson. *Finnish: A comprehensive grammar*. Routledge, 2017.

[76] Mark T Rivera, Sara B Soderstrom, and Brian Uzzi. “Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms”. In: *Annual Review of Sociology* 36 (2010), pp. 91–115.

[77] Aili Asikainen et al. “Cumulative effects of triadic closure and homophily in social networks”. In: *Science Advances* 6.19 (2020), eaax7310.

[78] Mark S Granovetter. “The strength of weak ties”. In: *American journal of sociology* 78.6 (1973), pp. 1360–1380.

[79] Maksim Kitsak. “Latent geometry for complementarity-driven networks”. In: *arXiv preprint arXiv:2003.06665* (2020).

[80] István A Kovács et al. “Network-based prediction of protein interactions”. In: *Nature communications* 10.1 (2019), pp. 1–8.

[81] Carolina ES Mattsson et al. “Functional structure in production networks”. In: *Frontiers in big Data* 4 (2021), p. 666712.

[82] Mingshan Jia, Bogdan Gabrys, and Katarzyna Musial. “Measuring quadrangle formation in complex networks”. In: *IEEE Transactions on Network Science and Engineering* 9.2 (2021), pp. 538–551.

[83] Hao Yin, Austin R Benson, and Jure Leskovec. “The local closure coefficient: A new perspective on network clustering”. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019, pp. 303–311.

[84] James HY Tai. “Chinese classifier systems and human categorization”. In: *In honor of William S.-Y. Wang: Interdisciplinary studies on language and language change* (1994), pp. 479–494.

[85] Lisa L-S Cheng and Rint Sybesma. “Yi-wan tang, yi-ge tang: Classifiers and massifiers”. In: *Tsing Hua journal of Chinese studies* 28.3 (1998), pp. 385–412.

[86] Ethan White, Brian Enquist, and Jessica Green. “On estimating the exponent of power-law frequency distributions”. In: *Ecology* 89 (May 2008), pp. 905–12. DOI: `10.1890/07-1288.1`.

[87] Nicolò Vallarano et al. “Fast and scalable likelihood maximization for Exponential Random Graph Models with local constraints”. In: *Scientific Reports* 11.1 (2021), pp. 1–33.

[88] Tiziano Squartini, Rossana Mastrandrea, and Diego Garlaschelli. “Unbiased sampling of network ensembles”. In: *New Journal of Physics* 17.2 (2015), p. 023052.

A

Appendix

This appendix presents the method of logarithmic binning used across the thesis.

In the linear binning, every bin has the same linear size $(k_{i+1} - k_i)$, while in the logarithmic binning, bins have constant logarithmic width b, where $b = \log(k_{i+1}) - \log(k_i)$ [86]. Thus, the linear bin width of a logarithmically-binned bin, $w_i = k_{i+1} - k_i = k_i(e^b - 1)$, is proportional to k_i. The sizes of logarithmic bins grow exponentially. Therefore, the number of observations x in a bin is equal to the density of observations $f(k)$ in that bin times the width w of that bin.

A.1. Power-law Exponent Estimation

The above mentioned method is utilized for most analysis. However, we need extra steps to estimate the power-law exponents. Since the probability density function $f(k)$ is proportional to $k^{-\gamma}$, the number of observations $x \propto f(k) \times w \propto k^{1-\gamma}$. Regressing $\log(x)$ against $\log(k)$ yields a slope equal to $1 - \gamma$.

To estimate γ accurately, the normalization of number of observations x is required. Due to the linearly increasing width of bins, a bin can contain more than one value of degree value k. The sum of all observations within a bin is x. To preserve the probability of a node with degree k such that the total probability of degree distribution is equal to 1, the number of observations x should be normalized by the linear width of the bin. This converts x to the number of observations per unit of bin, so $(x/w) \propto k^{-\gamma}$. As a result, regressing the normalized logarithmic bin counts $\log(x/w)$ against the logarithmic degree $\log(k)$ yields a slope of $-\gamma$ [86].

Additionally, the choice of the number of bins matters, since we want to decrease the number of empty and low-count bins to get a better estimation of γ [86]. Thus, according to different sizes of networks, the bin numbers vary. However, finding the optimal number of bins for each network requires future work.

B

Appendix

This appendix presents the complete statistics of semantic networks in ten languages. For each language, the statistics are computed for both the full network and the largest connected component.

B.1. French

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	27,598	5,254	1,405,857	1,411,997	12,293	68,591
L	31,099	5,512	2,321,424	2,346,989	8,649	63,080
d_{max}	773	804	78952	78957	49	113
$E[D]$	2.25	2.10	3.30	3.32	1.41	1.84
ANND	49.24	101.09	1371.08	1363.67	2.11	3.89
ANND rewired	39.92	81.55	2815.22	2791.08	2.38	4.34
ANND reconstructed	39.91	73.05	2572.11	2551.63	2.40	4.37
c_G	0.0921176	0.0709545	0.0725957	0.0731421	0.0122623	0.1145248
c_G *rewired*	0.0040955	0.0151831	0.0057348	0.0056810	0.0000322	0.0000257
c_G *reconstructed*	0.0036711	0.0159415	0.0060876	0.0060172	0.0000843	0.0000186
$\check{c}_G$	0.0103522	0.0046753	0.0004334	0.0004503	0.0237738	0.1797682
$\check{c}_G$ *rewired*	0.0038628	0.0028443	0.0003672	0.0003686	0.0002503	0.0000850
$\check{c}_G$ *reconstructed*	0.0045368	0.0034957	0.0004687	0.0004716	0.0002503	0.0001133
γ	2.5	2.4	2.3	2.3	3.4	3.3

Table B.1: Statistics of six French semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	17,519	2,832	1,289,083	1,296,622	1,361	20,144
L	23,157	3,549	2,216,411	2,243,044	1,665	28,264
d_{max}	773	804	78952	78957	49	113
$E[D]$	2.64	2.51	3.44	3.46	2.45	2.81
ANND	73.58	183.29	1492.49	1482.26	5.99	8.10
ANND rewired	54.48	144.75	3012.17	2976.66	5.23	6.97
ANND reconstructed	51.69	112.70	2717.38	2686.65	5.40	7.03
c_G	0.1330360	0.1177744	0.0790695	0.0795575	0.0084102	0.1563091
c_G *rewired*	0.0090988	0.0419280	0.0068039	0.0067468	0.0019302	0.0003799
c_G *reconstructed*	0.0111658	0.0438683	0.0072894	0.0072133	0.0030997	0.0003367
$\check{c}_G$	0.0102118	0.0044024	0.0004335	0.0004504	0.0067114	0.1516492
$\check{c}_G$ *rewired*	0.0060552	0.0042900	0.0003859	0.0003939	0.0046141	0.0006007
$\check{c}_G$ *reconstructed*	0.0086643	0.0061726	0.0004980	0.0005021	0.0058725	0.0007773
γ	2.4	2.3	2.3	2.3	2.7	3.1

Table B.2: Statistics of largest connected component of six French semantic networks extracted from ConceptNet.

B.2. Italian

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	3,816	45	108,983	110,770	2,847	21,928
L	4,675	38	98,743	103,343	1,611	16,038
d_{max}	247	6	1355	1355	7	28
$E[D]$	2.45	1.69	1.81	1.87	1.13	1.46
ANND	41.83	2.45	17.36	18.54	1.26	1.98
ANND rewired	38.84	2.66	16.04	17.19	1.32	2.44
ANND reconstructed	36.85	2.45	15.83	16.86	1.33	2.45
c_G	0.0632934	0.0000000	0.0579386	0.0582161	0.0000000	0.0685227
c_G *rewired*	0.0262871	0.0162963	0.0002225	0.0002295	0.0000000	0.0000181
c_G *reconstructed*	0.0261353	0.0459259	0.0001485	0.0002521	0.0000000	0.0000184
$\check{c}_G$	0.0073074	0.0000000	0.0139081	0.0133864	0.0000000	0.2991435
$\check{c}_G$ *rewired*	0.0128538	0.0526316	0.0002717	0.0003603	0.0000000	0.0001298
$\check{c}_G$ *reconstructed*	0.0181390	0.0526316	0.0002717	0.0004026	0.0000000	0.0001298
γ	2.4	None	2.8	2.8	4.6	3.8

Table B.3: Statistics of six Italian semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	2,663	9	36,295	46,468	13	1,580
L	3,802	13	39,961	52,729	12	2,010
d_{max}	247	6	1355	1355	7	15
$E[D]$	2.86	2.89	2.20	2.27	1.85	2.54
ANND	58.03	4.63	44.35	39.19	3.82	4.31
ANND rewired	48.54	4.52	36.68	32.66	3.86	4.31
ANND reconstructed	44.36	3.47	35.87	32.21	3.62	4.23
c_G	0.0847080	0.0000000	0.0533138	0.0601817	0.0000000	0.1424910
c_G *rewired*	0.0553913	0.6888889	0.0018624	0.0012401	0.1062271	0.0022310
c_G *reconstructed*	0.0597332	0.3259259	0.0015406	0.0012814	0.0000000	0.0009228
$\check{c}_G$	0.0070217	0.0000000	0.0065990	0.0085331	0.0000000	0.2774144
$\check{c}_G$ *rewired*	0.0181617	0.3947368	0.0017431	0.0014382	0.1111111	0.0041065
$\check{c}_G$ *reconstructed*	0.0273120	0.3214286	0.0019928	0.0015799	0.0000000	0.0031939
γ	2.3	None	2.6	2.6	None	3.7

Table B.4: Statistics of largest connected component of six Italian semantic networks extracted from ConceptNet.

B.3. German

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	122,810	50	108,479	176,164	3,545	80,271
L	162,971	34	245,089	392,221	2,691	103,967
d_{max}	2228	4	2542	2593	16	284
$E[D]$	2.65	1.36	4.52	4.45	1.52	2.59
ANND	80.17	2.12	109.23	100.43	2.17	8.68
ANND rewired	50.12	1.77	79.98	82.09	2.42	9.79
ANND reconstructed	48.59	1.72	78.66	80.51	2.44	9.87
c_G	0.0656267	0.0000000	0.0812345	0.0870983	0.0147656	0.1419571
c_G *rewired*	0.0014066	0.0000000	0.0039561	0.0025991	0.0000000	0.0001979
c_G *reconstructed*	0.0014423	0.0000000	0.0040200	0.0028831	0.0000000	0.0001646
$\check{c}_G$	0.0092784	0.0000000	0.0078443	0.0110564	0.0377854	0.1040466
$\check{c}_G$ *rewired*	0.0026754	0.0000000	0.0026620	0.0030590	0.0000000	0.0003251
$\check{c}_G$ *reconstructed*	0.0031767	0.0000000	0.0031971	0.0036737	0.0000000	0.0003022
γ	2.5	None	2.6	2.5	3.6	3.1

Table B.5: Statistics of six German semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	113,301	5	100,737	172,147	187	43,072
L	155,806	4	240,482	389,847	202	76,898
d_{max}	2228	4	2542	2593	10	284
$E[D]$	2.75	1.60	4.77	4.53	2.16	3.57
ANND	86.42	3.40	117.51	102.74	3.89	13.85
ANND rewired	52.56	3.40	82.41	82.39	3.50	12.50
ANND reconstructed	51.49	1.40	80.49	81.55	3.27	12.42
c_G	0.0705959	0.0000000	0.0859303	0.0887972	0.0000000	0.1775439
c_G *rewired*	0.0015456	0.0000000	0.0042611	0.0027321	0.0051821	0.0004686
c_G *reconstructed*	0.0017990	0.0000000	0.0043755	0.0029790	0.0108480	0.0004635
$\check{c}_G$	0.0092892	0.0000000	0.0078301	0.0110530	0.0000000	0.0966821
$\check{c}_G$ *rewired*	0.0029596	0.0000000	0.0026440	0.0031052	0.0125261	0.0006826
$\check{c}_G$ *reconstructed*	0.0035600	0.0000000	0.0033392	0.0036694	0.0187891	0.0007412
γ	2.5	None	2.6	2.5	None	3.1

Table B.6: Statistics of largest connected component of six German semantic networks extracted from ConceptNet.

B.4. Spanish

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	1,253	81	104,445	105,053	1,414	14,437
L	1,199	64	101,354	102,593	836	12,663
d_{max}	70	6	65	76	7	61
$E[D]$	1.91	1.58	1.94	1.95	1.18	1.75
ANND	14.02	2.80	41.53	41.45	1.40	2.86
ANND rewired	12.46	2.57	22.47	22.42	1.42	3.17
ANND reconstructed	11.57	2.37	22.42	22.36	1.44	3.20
c_G	0.0214615	0.0000000	0.0284659	0.0284010	0.0000000	0.1320876
c_G *rewired*	0.0126838	0.0000000	0.0002006	0.0002151	0.0000000	0.0000000
c_G *reconstructed*	0.0115027	0.0000000	0.0001852	0.0001602	0.0000000	0.0001386
$\check{c}_G$	0.0138260	0.0000000	0.0042407	0.0043782	0.0000000	0.2585298
$\check{c}_G$ *rewired*	0.0160481	0.0000000	0.0022488	0.0021946	0.0000000	0.0000000
$\check{c}_G$ *reconstructed*	0.0172825	0.0000000	0.0022184	0.0022315	0.0000000	0.0003229
γ	2.6	None	2.6	2.4	4.4	3.6

Table B.7: Statistics of six Spanish semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	255	11	12,094	22,861	15	3,491
L	312	15	12,858	24,392	14	4,481
d_{max}	32	6	65	76	4	61
$E[D]$	2.45	2.73	2.13	2.13	1.87	2.57
ANND	8.91	4.27	42.53	41.49	2.89	5.76
ANND rewired	8.01	3.97	22.57	21.90	2.71	5.07
ANND reconstructed	7.82	2.89	22.45	21.65	2.68	5.02
c_G	0.0163449	0.0000000	0.0414353	0.0402895	0.0000000	0.1637785
c_G *rewired*	0.0270048	0.3393939	0.0025752	0.0010344	0.0000000	0.0017602
c_G *reconstructed*	0.0196794	0.3939394	0.0025220	0.0014533	0.0000000	0.0014583
$\check{c}_G$	0.0154719	0.0000000	0.0083946	0.0080899	0.0000000	0.1813969
$\check{c}_G$ *rewired*	0.0355854	0.2926829	0.0147184	0.0080720	0.0000000	0.0026171
$\check{c}_G$ *reconstructed*	0.0355854	0.3913043	0.0153530	0.0085750	0.0000000	0.0027807
γ	None	None	2.5	2.3	None	3.0

Table B.8: Statistics of largest connected component of six Spanish semantic networks extracted from ConceptNet.

B.5. Russian

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	1,748	26	56,978	57,570	1,857	12,582
L	1,622	15	83,705	85,286	1,079	9,501
d_{max}	105	2	113	117	7	31
$E[D]$	1.86	1.15	2.94	2.96	1.16	1.51
ANND	13.20	1.31	8.70	9.17	1.36	2.41
ANND rewired	8.96	1.27	8.13	8.34	1.39	2.50
ANND reconstructed	9.14	1.27	8.08	8.31	1.38	2.53
c_G	0.0273367	0.0000000	0.1758516	0.1729486	0.0026925	0.0597137
c_G *rewired*	0.0040824	0.0000000	0.0000930	0.0000816	0.0000000	0.0000000
c_G *reconstructed*	0.0037436	0.0000000	0.0001221	0.0001240	0.0000000	0.0000000
$\check{c}_G$	0.0130687	0.0000000	0.2038037	0.1937980	0.0432692	0.1360682
$\check{c}_G$ *rewired*	0.0070184	0.0000000	0.0003143	0.0002591	0.0000000	0.0000000
$\check{c}_G$ *reconstructed*	0.0055663	0.0000000	0.0003345	0.0002879	0.0000000	0.0000000
γ	2.5	None	2.9	2.9	4.4	3.6

Table B.9: Statistics of six Russian semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	557	3	20,268	25,887	12	1,148
L	622	2	41,981	50,262	11	1,299
d_{max}	105	2	113	117	7	23
$E[D]$	2.23	1.33	4.14	3.88	1.83	2.26
ANND	24.80	1.67	11.54	12.05	5.16	5.35
ANND rewired	16.69	1.67	10.51	10.30	4.57	4.40
ANND reconstructed	15.46	1.67	10.39	10.32	2.19	4.35
c_G	0.0419918	0.0000000	0.2436609	0.2236226	0.0000000	0.0952778
c_G *rewired*	0.0246127	0.0000000	0.0007270	0.0006514	0.1011905	0.0018042
c_G *reconstructed*	0.0252673	0.0000000	0.0005522	0.0004168	0.0000000	0.0011219
$\check{c}_G$	0.0115815	0.0000000	0.2128178	0.1965126	0.0000000	0.1033279
$\check{c}_G$ *rewired*	0.0135118	0.0000000	0.0011607	0.0008966	0.1071429	0.0041890
$\check{c}_G$ *reconstructed*	0.0239028	0.0000000	0.0009331	0.0008902	0.0000000	0.0041890
γ	None	None	3.6	3.5	None	3.0

Table B.10: Statistics of largest connected component of six Russian semantic networks extracted from ConceptNet.

B.6. Portuguese

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	5,667	146	34,541	37,760	1,427	18,299
L	5,189	108	33,593	38,862	894	17,142
d_{max}	110	8	294	294	6	34
$E[D]$	1.83	1.48	1.95	2.06	1.25	1.87
ANND	9.37	2.87	35.01	30.88	1.53	2.73
ANND rewired	8.48	2.51	21.57	20.25	1.57	3.37
ANND reconstructed	8.21	2.51	21.43	20.29	1.56	3.37
c_G	0.0165339	0.0093770	0.0527061	0.0495699	0.0051390	0.1308075
c_G *rewired*	0.0012414	0.0076321	0.0010061	0.0009226	0.0000000	0.0000619
c_G *reconstructed*	0.0016592	0.0010274	0.0010767	0.0008424	0.0000000	0.0000000
$\check{c}_G$	0.0081459	0.0171429	0.0106169	0.0111971	0.0354331	0.2870057
$\check{c}_G$ *rewired*	0.0030347	0.0171429	0.0052403	0.0040381	0.0000000	0.0001470
$\check{c}_G$ *reconstructed*	0.0043125	0.0171429	0.0056623	0.0038392	0.0000000	0.0000000
γ	2.6	None	2.4	2.6	4.2	3.7

Table B.11: Statistics of six Portuguese semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	3,341	15	5,929	11,426	17	6,421
L	3,737	20	7,392	15,117	17	9,125
d_{max}	110	8	294	294	6	34
$E[D]$	2.24	2.67	2.49	2.65	2.00	2.84
ANND	14.75	5.52	51.62	25.73	3.35	4.67
ANND rewired	11.18	5.48	37.89	22.19	3.31	4.80
ANND reconstructed	11.15	3.45	35.79	22.42	3.39	4.73
c_G	0.0268333	0.0912698	0.1030854	0.0893813	0.0000000	0.2353418
c_G *rewired*	0.0030750	0.4050794	0.0160122	0.0048174	0.0000000	0.0004554
c_G *reconstructed*	0.0041104	0.3549206	0.0191889	0.0052573	0.0000000	0.0004310
$\check{c}_G$	0.0081021	0.0447761	0.0112864	0.0168496	0.0000000	0.2830883
$\check{c}_G$ *rewired*	0.0050741	0.2238806	0.0090722	0.0056969	0.0000000	0.0011397
$\check{c}_G$ *reconstructed*	0.0084295	0.3260870	0.0149447	0.0063813	0.0000000	0.0008767
γ	2.6	None	2.4	2.5	None	4.4

Table B.12: Statistics of largest connected component of six Portuguese semantic networks extracted from ConceptNet.

B.7. Dutch

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	1,278	142	19,859	20,790	2,127	21,983
L	1,244	194	16,550	17,943	1,378	28,720
d_{max}	40	24	27	45	17	34
$E[D]$	1.95	2.73	1.67	1.73	1.30	2.61
ANND	6.25	7.2	3.85	4.13	1.70	3.88
ANND rewired	7.62	10.41	3.08	3.67	1.74	4.89
ANND reconstructed	7.39	9.92	3.06	3.65	1.76	4.91
c_G	0.0330612	0.110967	0.0413025	0.0415835	0.0025074	0.2682013
c_G *rewired*	0.0032895	0.0454031	0.0000000	0.0000132	0.0000000	0.0002374
c_G *reconstructed*	0.0050336	0.0454412	0.0000108	0.0000716	0.0000000	0.0001786
$\check{c}_G$	0.1483806	0.234917	0.0650352	0.0977919	0.0057637	0.3276525
$\check{c}_G$ *rewired*	0.0131810	0.0866496	0.0000000	0.0001259	0.0000000	0.0002688
$\check{c}_G$ *reconstructed*	0.0139342	0.0997963	0.0000882	0.0004408	0.0000000	0.0003495
γ	2.3	None	4.1	3.0	4.4	3.9

Table B.13: Statistics of six Dutch semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	191	53	303	1,418	111	11,964
L	447	132	348	1,908	117	21,146
d_{max}	40	24	17	45	17	34
$E[D]$	4.68	4.98	2.3	2.69	2.11	3.53
ANND	16.54	14.61	5.02	7.31	5.12	5.62
ANND rewired	16.37	16.19	4.12	9.06	4	5.89
ANND reconstructed	14.37	12.77	4.34	8.47	3.99	5.85
c_G	0.115344	0.240704	0.0903214	0.0618426	0	0.3373450
c_G *rewired*	0.130678	0.317215	0.008369	0.0039301	0.0426288	0.0007485
c_G *reconstructed*	0.135231	0.274288	0.0053362	0.0052691	0.0101086	0.0004559
$\check{c}_G$	0.193433	0.246269	0.106857	0.1838883	0	0.3090331
$\check{c}_G$ *rewired*	0.136781	0.262551	0.0160285	0.0101128	0.0521739	0.0006107
$\check{c}_G$ *reconstructed*	0.173246	0.294923	0.0053428	0.0150660	0.0347826	0.0006107
γ	None	None	None	2.2	None	4.8

Table B.14: Statistics of largest connected component of six Dutch semantic networks extracted from ConceptNet.

B.8. Japanese

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	457	41,809	8,300	14,156	44,758	2,235	11,147
L	401	90,002	15,585	20,810	104,668	1,378	8,477
d_{max}	15	1118	225	145	1163	7	15
$E[D]$	1.75	4.31	3.76	2.94	4.68	1.23	1.52
ANND	3.07	87.79	23.36	12.36	86.75	1.44	2.16
ANND rewired	2.98	57.03	22.57	16.26	58.14	1.51	2.53
ANND reconstructed	2.91	56.30	21.45	16.02	57.14	1.51	2.51
c_G	0.0056285	0.1782808	0.1063128	0.1344127	0.1778910	0.0019985	0.0897062
c_G *rewired*	0.0001858	0.0062874	0.0060578	0.0024470	0.0064016	0.0000000	0.0000000
c_G *reconstructed*	0.0003456	0.0071730	0.0066768	0.0025707	0.0064782	0.0000000	0.0000000
$\check{c}_G$	0.0156454	0.0278480	0.0720310	0.1095622	0.0315147	0.0129870	0.3336179
$\check{c}_G$ *rewired*	0.0039113	0.0063782	0.0091209	0.0051749	0.0062536	0.0000000	0.0000000
$\check{c}_G$ *reconstructed*	0.0078227	0.0075674	0.0113028	0.0052705	0.0076118	0.0000000	0.0000000
γ	None	2.4	2.3	2.3	2.3	4.6	3.3

Table B.15: Statistics of seven Japanese semantic networks extracted from ConceptNet.

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	38	40,256	7,230	7,200	43,286	20	230
L	55	89,040	14,855	15,623	103,778	20	314
d_{max}	15	1118	225	145	1163	3	15
$E[D]$	2.89	4.42	4.11	4.34	4.79	2	2.73
ANND	8.12	91.11	26.54	22.19	89.64	2.42	4.98
ANND rewired	7.03	58.69	23.38	21.30	59.06	2.35	5.27
ANND reconstructed	6.71	57.00	22.74	21.02	57.80	2.37	5.54
c_G	0.0676901	0.1846882	0.1199764	0.1793375	0.1835516	0	0.187087
c_G *rewired*	0.142064	0.0072961	0.0088438	0.0065473	0.0068235	0.0833333	0.0277398
c_G *reconstructed*	0.165176	0.0074941	0.0081923	0.0075345	0.0074216	0	0.0147212
$\check{c}_G$	0.0446097	0.0278460	0.0720742	0.1055655	0.0315140	0	0.277652
$\check{c}_G$ *rewired*	0.111524	0.0062355	0.0102078	0.0109094	0.0064250	0.111111	0.0293454
$\check{c}_G$ *reconstructed*	0.130952	0.0075228	0.0128349	0.0113982	0.0080536	0	0.0293454
γ	None	2.4	2.3	2.2	2.3	None	None

Table B.16: Statistics of largest connected component of seven Japanese semantic networks extracted from ConceptNet.

B.9. Finnish

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	2,615	17	35,800	37,718	1,730	24,595
L	2,152	14	32,340	34,443	1,033	17,660
d_{max}	64	11	86	86	7	38
$E[D]$	1.65	1.65	1.81	1.83	1.19	1.44
ANND	8.35	7.59	14.77	14.72	1.42	2.04
ANND rewired	6.02	7.24	9.29	9.27	1.45	2.18
ANND reconstructed	5.94	2.13	9.29	9.30	1.47	2.18
c_G	0.0216604	0.0000000	0.0500543	0.0517182	0.0000000	0.0862996
c_G *rewired*	0.0004001	0.0000000	0.0000996	0.0002309	0.0000000	0.0000000
c_G *reconstructed*	0.0003591	0.0000000	0.0002141	0.0002135	0.0000000	0.0000000
$\check{c}_G$	0.0128492	0.0000000	0.0228548	0.0236020	0.0000000	0.2060870
$\check{c}_G$ *rewired*	0.0032123	0.0000000	0.0007734	0.0009631	0.0000000	0.0000000
$\check{c}_G$ *reconstructed*	0.0037964	0.0000000	0.0011097	0.0011536	0.0000000	0.0000000
γ	2.7	None	3.0	2.9	4.3	4.0

Table B.17: Statistics of six Finnish semantic networks extracted from ConceptNet.

Network	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	76	12	4,483	6,958	24	1,569
L	75	11	5,152	7,846	23	1,756
d_{max}	64	11	86	86	6	25
$E[D]$	1.97	1.83	2.30	2.26	1.92	2.24
ANND	54.32	10.17	18.90	19.57	3.41	4.21
ANND rewired	52.96	10.17	10.97	10.95	2.96	3.75
ANND reconstructed	20.80	2.67	10.98	11.12	2.71	3.69
c_G	0.0000000	0.0000000	0.0623683	0.0580776	0.0000000	0.0891202
c_G *rewired*	0.0000000	0.0000000	0.0021050	0.0014318	0.0513889	0.0000000
c_G *reconstructed*	0.0026484	0.0000000	0.0030546	0.0017647	0.1000000	0.0003612
$\check{c}_G$	0.0000000	0.0000000	0.0294814	0.0247665	0.0000000	0.0932416
$\check{c}_G$ *rewired*	0.0000000	0.0000000	0.0083471	0.0069679	0.0714286	0.0000000
$\check{c}_G$ *reconstructed*	0.0037453	0.0000000	0.0099455	0.0059455	0.1428571	0.0006258
γ	None	None	2.4	2.6	None	4.2

Table B.18: Statistics of largest connected component of six Finnish semantic networks extracted from ConceptNet.

B.10. Chinese

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	6,932	11,479	4,004	7,507	18,426	79	145,262
L	11,774	16,136	6,131	10,270	33,186	45	79,220
d_{max}	207	434	155	1129	435	3	10
$E[D]$	3.40	2.81	3.06	2.74	3.60	1.14	1.09
ANND	26.25	63.70	12.21	162.93	57.97	1.24	1.10
ANND rewired	22.44	41.65	13.98	185.39	42.40	1.23	1.25
ANND reconstructed	21.92	41.00	14.73	163.25	41.54	1.28	1.25
c_G	0.0179595	0.0297284	0.0358902	0.0083528	0.0369246	0	0.0489128
c_G *rewired*	0.0056755	0.0089165	0.0052837	0.0616974	0.0072875	0	0.0000000
c_G *reconstructed*	0.0069028	0.0098144	0.0039351	0.0665837	0.0072398	0	0.0000000
$\check{c}_G$	0.0174370	0.0127920	0.0579787	0.0003615	0.0232420	0	0.8347986
$\check{c}_G$ *rewired*	0.0119444	0.0150729	0.0107718	0.0032367	0.0139189	0	0.0000000
$\check{c}_G$ *reconstructed*	0.0146595	0.0192065	0.0083151	0.0049678	0.0162964	0	0.0000000
γ	2.5	2.3	2.7	1.9	2.3	None	5.6

Table B.19: Statistics of seven Chinese semantic networks extracted from ConceptNet.

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
N	6,355	10,073	3,417	3,163	17,128	4	17
L	11,366	15,212	5,748	6,424	32,344	3	19
d_{max}	207	434	155	1129	435	3	9
$E[D]$	3.58	3.02	3.36	4.06	3.78	1.5	2.24
ANND	28.38	72.27	13.96	383.10	62.20	2.5	3.79
ANND rewired	22.82	46.03	15.19	381.71	44.52	2.5	4.47
ANND reconstructed	23.00	43.50	15.21	212.58	42.50	1	4.25
c_G	0.0194065	0.0338779	0.0420557	0.0130727	0.0397228	0	0.0702614
c_G *rewired*	0.0073810	0.0118020	0.0065544	0.3140124	0.0089426	0	0.119281
c_G *reconstructed*	0.0084757	0.0131889	0.0065041	0.2059449	0.0092014	0	0.140523
$\check{c}_G$	0.0174636	0.0128126	0.0582090	0.0003313	0.0232570	0	0.0576923
$\check{c}_G$ *rewired*	0.0152167	0.0177354	0.0126360	0.0092905	0.0145521	0	0.0576923
$\check{c}_G$ *reconstructed*	0.0157160	0.0228107	0.0132431	0.0142171	0.0175878	0	0.115385
γ	2.5	2.3	2.7	1.9	2.3	None	None

Table B.20: Statistics of lcc of seven Chinese semantic networks extracted from ConceptNet.

C
Appendix

This Appendix shows the statistics of 4 properties of semantic networks from the eleven languages. Each property is compared among the seven networks for the eleven languages.

C.1. Maximum Degree

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
English	198	2,913	116	4,025	5,263	38	103
French		773	804	78,952	78,957	49	113
Italian		247	6	1,355	1,355	7	15
German		2,228	4	2,542	2,593	10	284
Spanish		32	6	65	76	4	61
Russian		105	2	113	117	7	23
Portuguese		110	8	294	294	6	34
Dutch		40	24	17	45	17	34
Japanese	15	1,118	225	145	1,163	3	15
Finnish		64	11	86	86	6	25
Chinese	207	434	155	1,129	435	3	9

Table C.1: Maximum degree d_{max} of semantic networks in different languages.

C.2. Degree Correlation Coefficient

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
English	-0.146332	-0.0827203	-0.155642	-0.0884337	-0.0837246	-0.0046235	0.104139
French		-0.120703	-0.14666	0.0024895	0.0023905	-0.100085	-0.0518867
Italian		-0.274367	-0.822804	-0.0333226	-0.0327359	-0.414141	0.0686239
German		-0.0508463	-1	-0.054565	-0.0602081	-0.303106	-0.05007
Spanish		-0.304536	-0.692961	-0.668534	-0.650121	-0.641026	-0.0616795
Russian		-0.18001	-1	-0.0561318	-0.0791576	-0.844636	-0.234003
Portuguese		-0.15915	-0.609884	-0.200251	-0.11843	-0.536266	0.0918521
Dutch		-0.357844	-0.531119	-0.227659	0.24378	-0.310662	0.118875
Japanese	-0.53168	-0.10287	-0.0883566	-0.0919621	-0.0987381	-0.212121	0.167217
Finnish		-0.783611	-1	-0.442636	-0.472006	-0.564626	-0.143689
Chinese	-0.0836939	-0.204188	-0.0057228	-0.268103	-0.138984	-1	-0.249394

Table C.2: Degree correlation coefficient ρ_D of semantic networks in different languages.

C.3. Average Clustering Coefficient

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
English	0.0021657	0.056611	0.0460611	0.102285	0.103635	0.0150401	0.112972
French		0.133036	0.117774	0.0790695	0.0795575	0.0084102	0.156309
Italian		0.084708	0	0.0533138	0.0601817	0	0.142491
German		0.0705959	0	0.0859303	0.0887972	0	0.177544
Spanish		0.0163449	0	0.0414353	0.0402895	0	0.163778
Russian		0.0419918	0	0.243661	0.223623	0	0.0952778
Portuguese		0.0268333	0.0912698	0.103085	0.0893813	0	0.235342
Dutch		0.115344	0.240704	0.0903214	0.0618426	0	0.337345
Japanese	0.0676901	0.184688	0.119976	0.179338	0.183552	0	0.187087
Finnish		0	0	0.0623683	0.0580776	0	0.0891202
Chinese	0.0194065	0.0338779	0.0420557	0.0130727	0.0397228	0	0.0702614

Table C.3: Average clustering coefficient c_G of seven semantic networks in different languages.

C.4. Graph Transitivity

Network	*Has-A*	*Is-A*	*Part-Of*	*Related-To*	*Union*	*Antonym*	*Synonym*
English	0.0011555	0.0021951	0.0179152	0.0080132	0.0072214	0.0218321	0.0907131
French		0.0102118	0.0044024	0.0004335	0.0004504	0.0067114	0.151649
Italian		0.0070217	0	0.006599	0.0085331	0	0.277414
German		0.0092892	0	0.0078301	0.011053	0	0.0966821
Spanish		0.0154719	0	0.0083946	0.0080899	0	0.181397
Russian		0.0115815	0	0.212818	0.196513	0	0.103328
Portuguese		0.0081021	0.0447761	0.0112864	0.0168496	0	0.283088
Dutch		0.193433	0.246269	0.106857	0.183888	0	0.309033
Japanese	0.0446097	0.027846	0.0720742	0.105566	0.031514	0	0.277652
Finnish		0	0	0.0294814	0.0247665	0	0.0932416
Chinese	0.0174636	0.0128126	0.058209	0.0003313	0.023257	0	0.0576923

Table C.4: Graph transitivity $\check{c}_G$ of seven semantic networks in different languages.

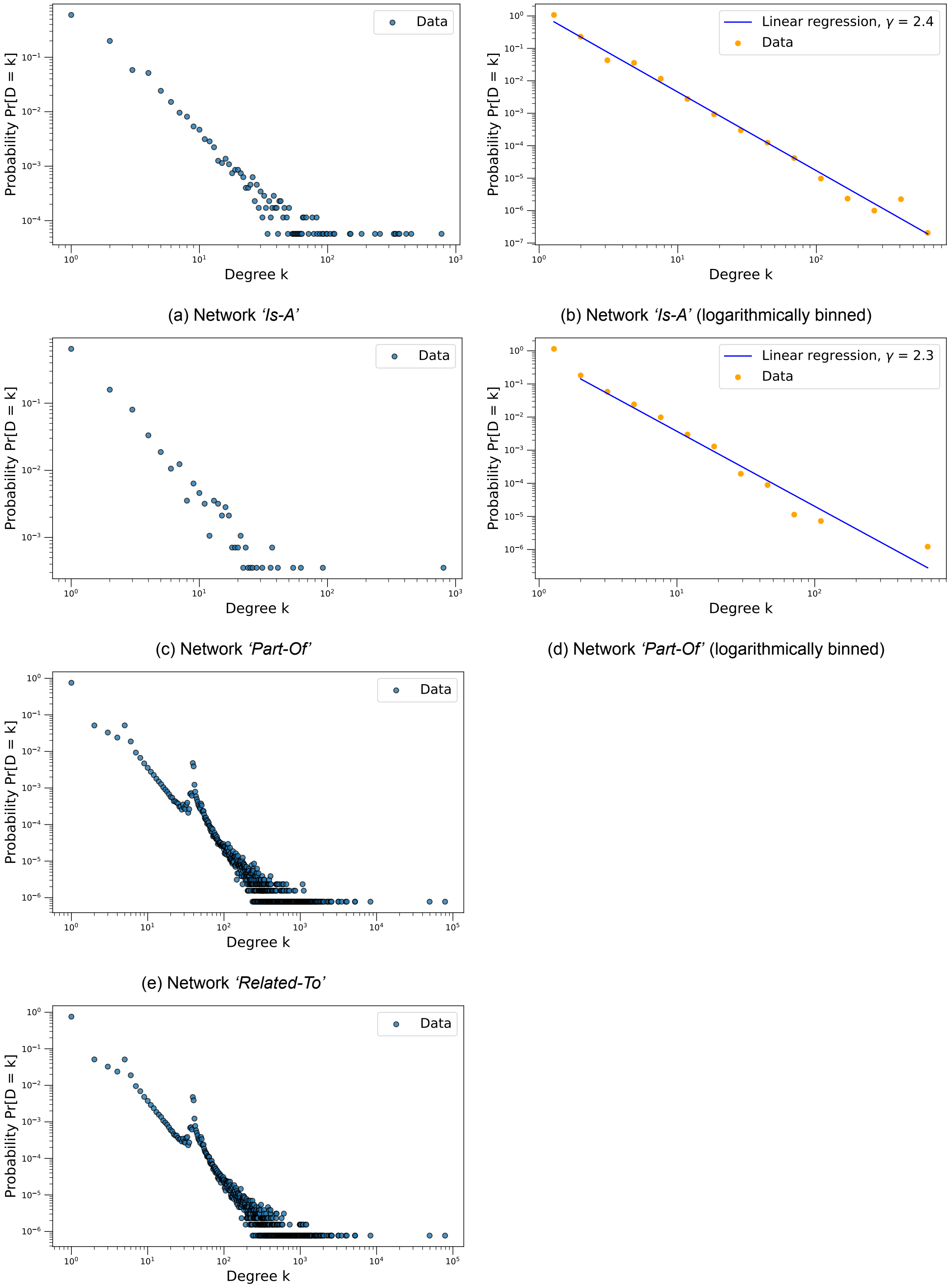
D

Appendix

This appendix provides the plots (in log-log scale) of degree distribution of all semantic networks, and the estimation of power-law exponents γ using logarithmic binning.

For networks have fewer than 1000 nodes, we do not show the degree distribution as we do not estimate their power-law exponent. Additionally, we do not estimate the power-law exponents γ for networks that do not have power-law degree distribution.

D.1. French



(a) Network *'Is-A'*

(b) Network *'Is-A'* (logarithmically binned)

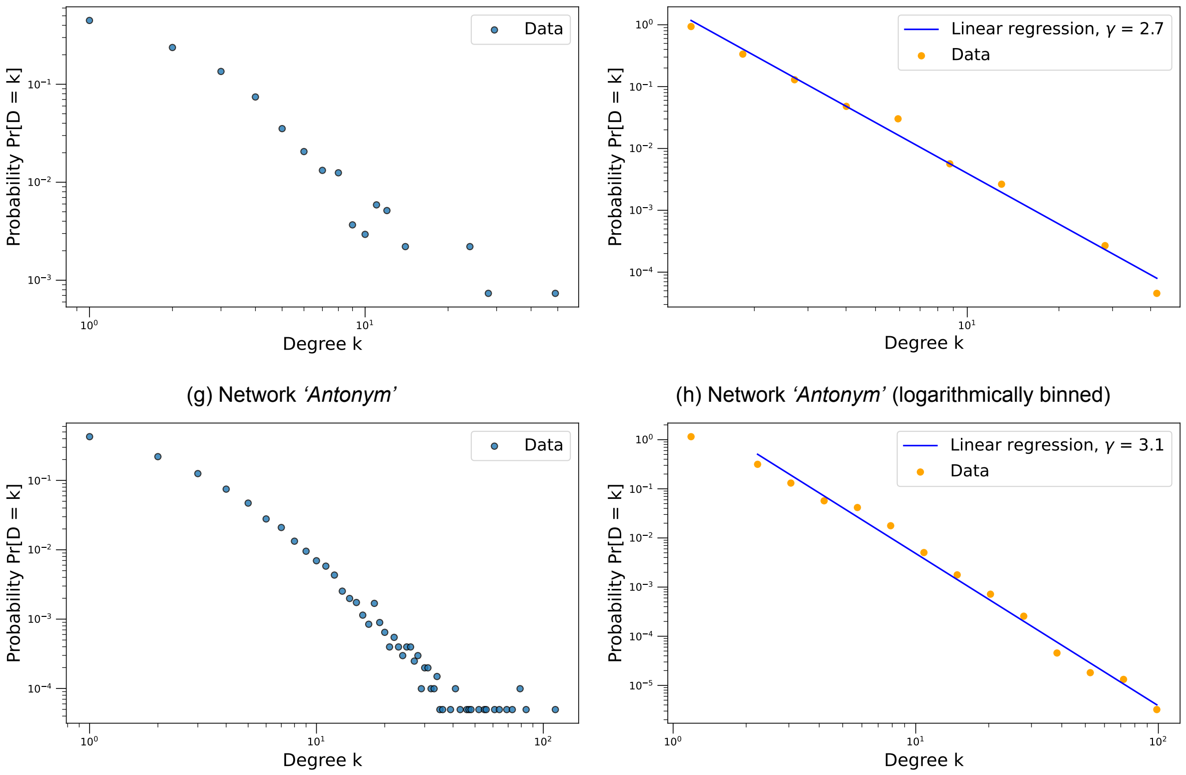
(c) Network *'Part-Of'*

(d) Network *'Part-Of'* (logarithmically binned)

(e) Network *'Related-To'*

(f) Network *'Union'*

Figure D.1: Degree distributions of six French semantic networks and power-law exponents estimation over logarithmically binned degree distribution.



(g) Network *'Antonym'*

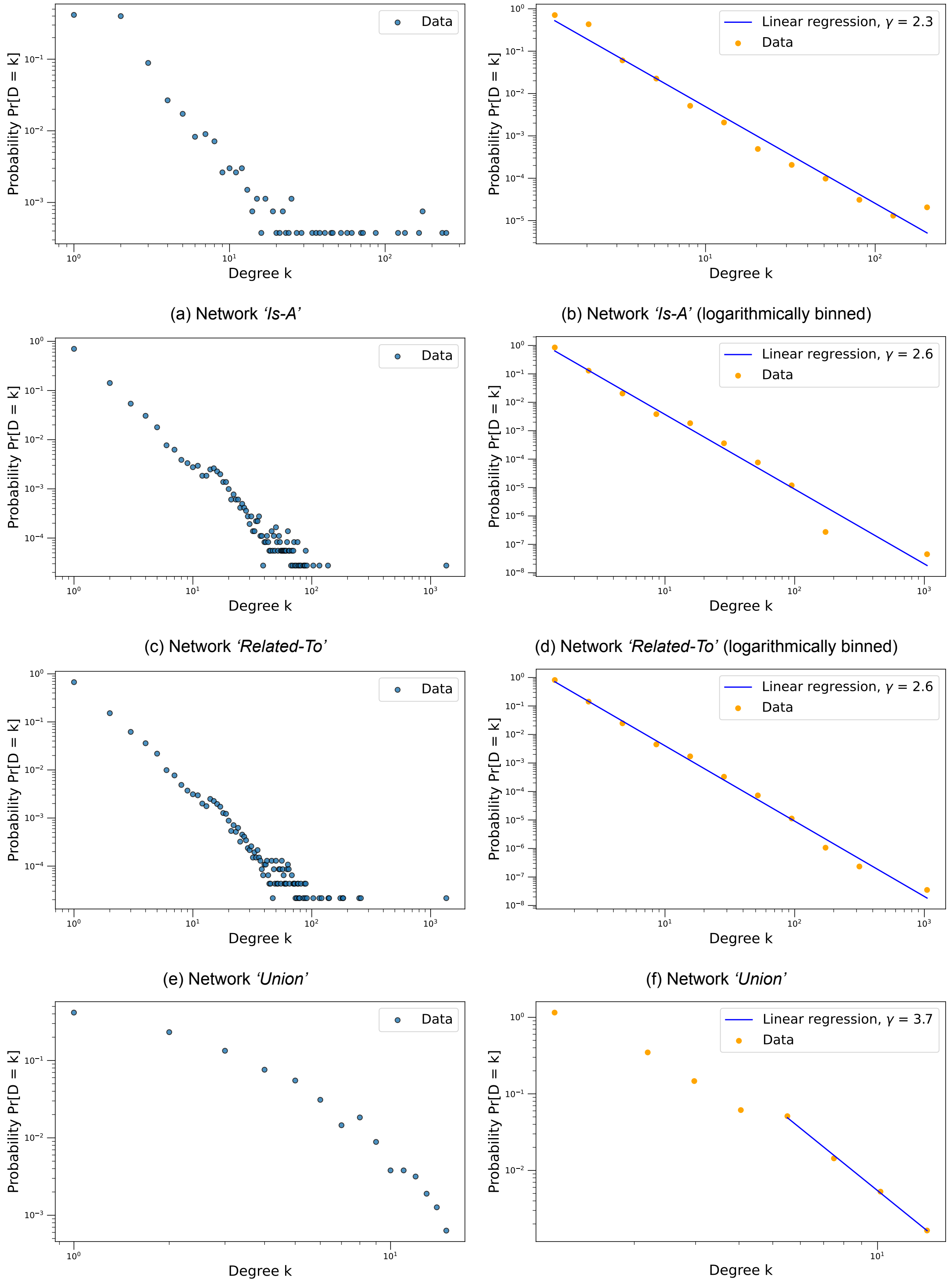
(h) Network *'Antonym'* (logarithmically binned)

(i) Network *'Synonym'*

(j) Network *'Synonym'* (logarithmically binned)

Figure D.1: Degree distributions of six French semantic networks and power-law exponents estimation over logarithmically binned degree distribution (cont.).

D.2. Italian



(a) Network *'Is-A'*

(b) Network *'Is-A'* (logarithmically binned)

(c) Network *'Related-To'*

(d) Network *'Related-To'* (logarithmically binned)

(e) Network *'Union'*

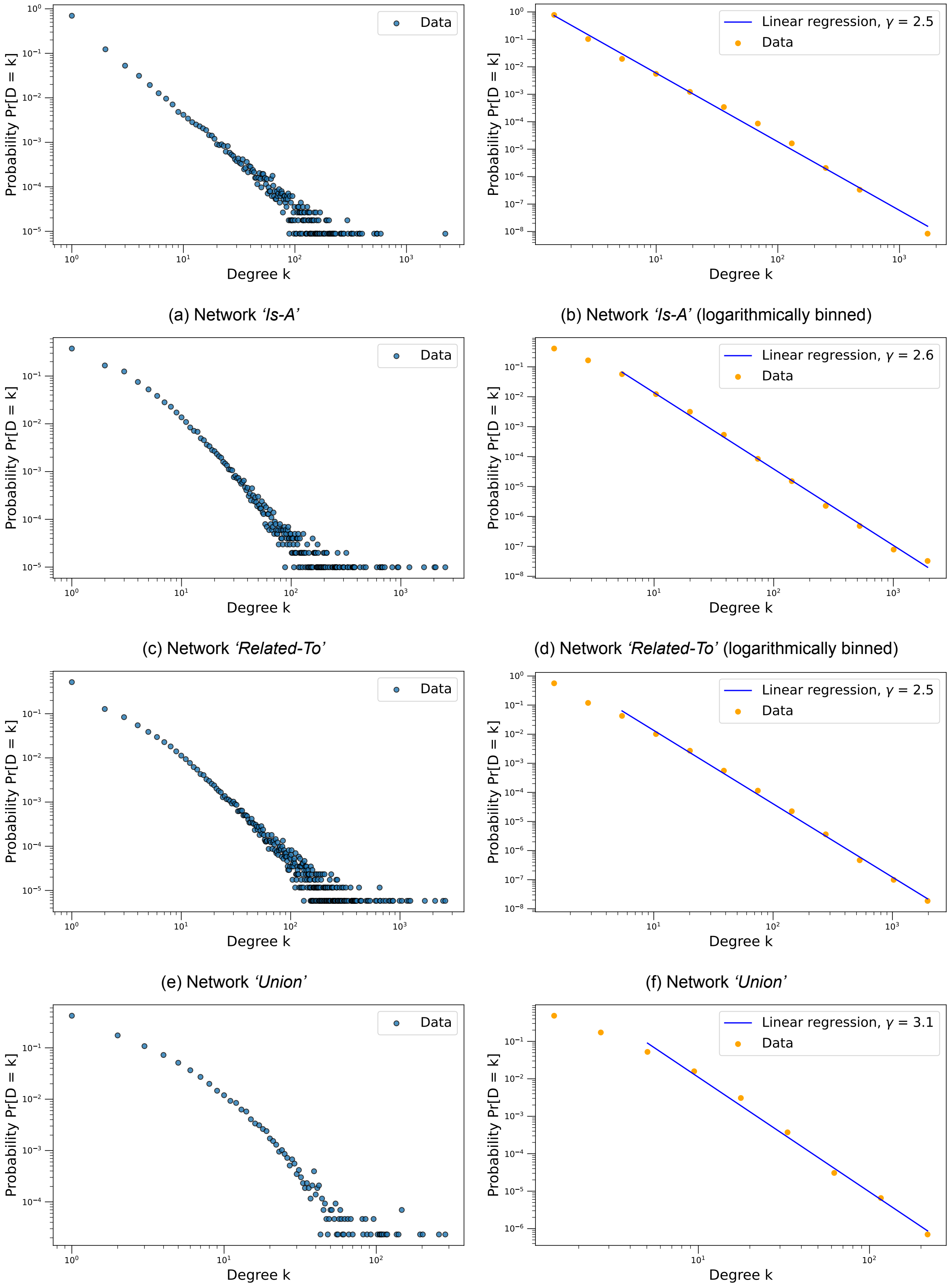
(f) Network *'Union'*

(g) Network *'Synonym'*

(h) Network *'Synonym'*

Figure D.2: Degree distributions of four Italian semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.3. German



(a) Network *‘Is-A’*

(b) Network *‘Is-A’* (logarithmically binned)

(c) Network *‘Related-To’*

(d) Network *‘Related-To’* (logarithmically binned)

(e) Network *‘Union’*

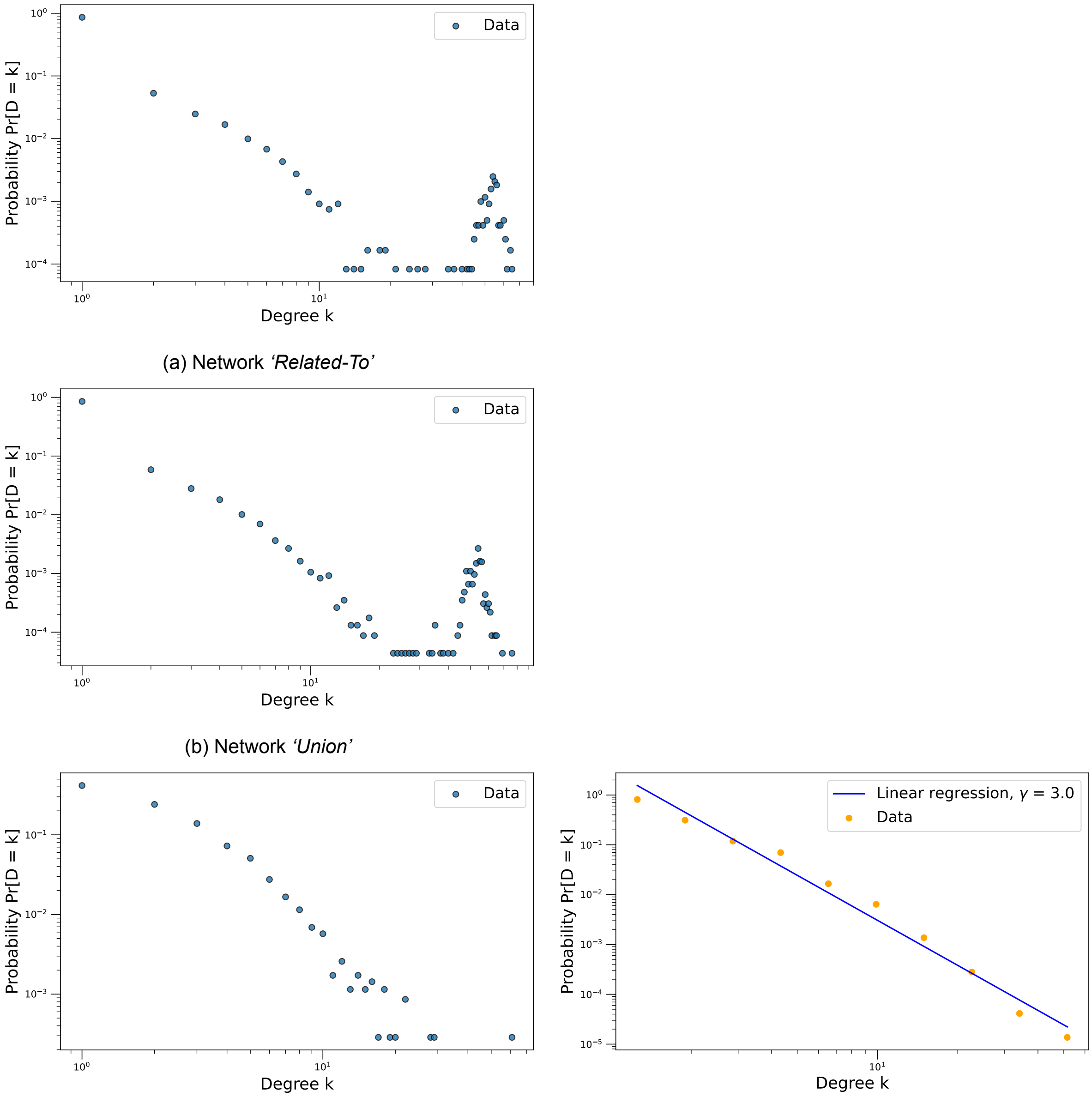
(f) Network *‘Union’*

(g) Network *‘Synonym’*

(h) Network *‘Synonym’*

Figure D.3: Degree distributions of four German semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.4. Spanish



(a) Network *'Related-To'*

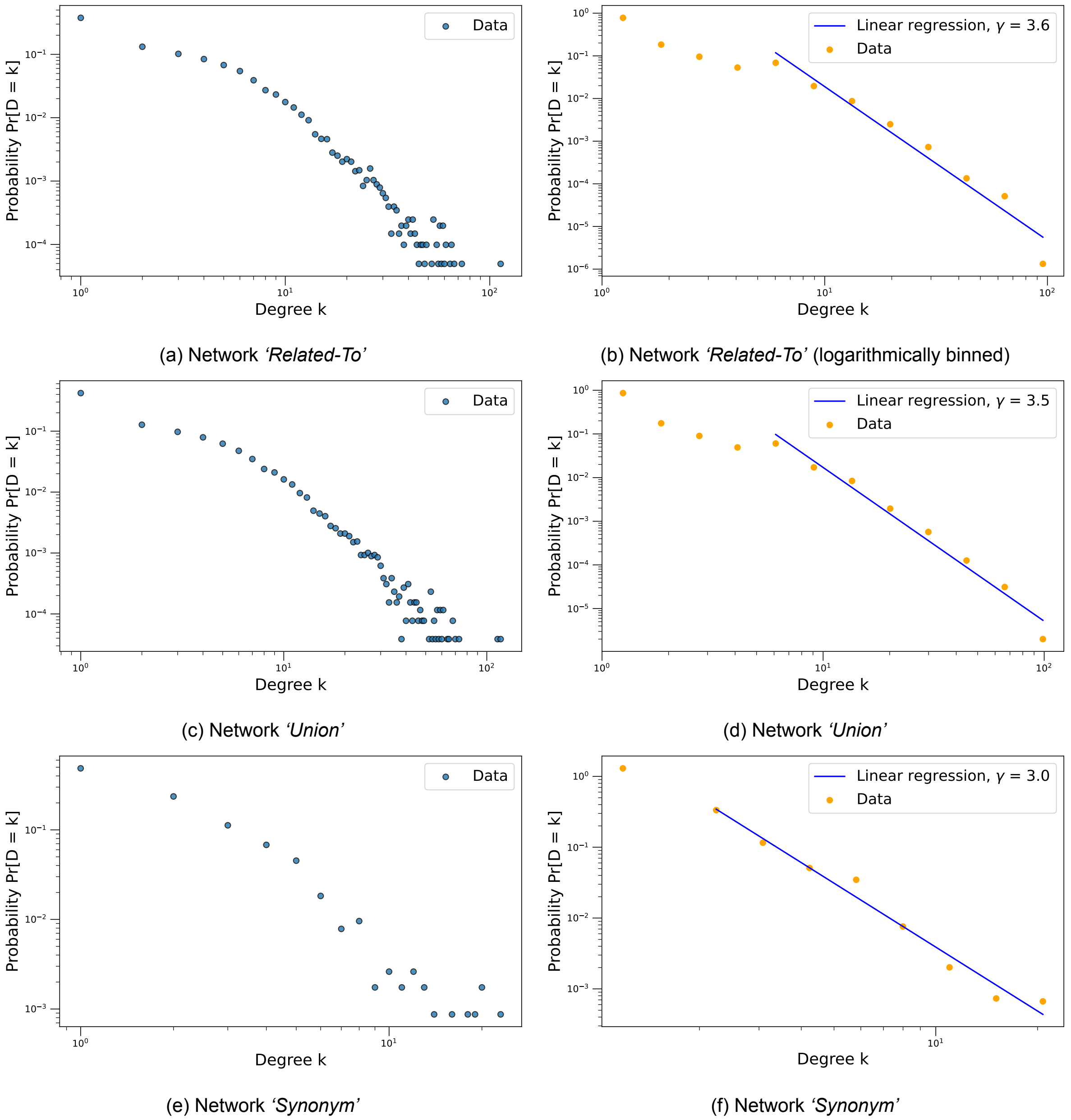
(b) Network *'Union'*

(c) Network *'Synonym'*

(d) Network *'Synonym'*

Figure D.4: Degree distributions of three Spanish semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.5. Russian



(a) Network *'Related-To'*

(b) Network *'Related-To'* (logarithmically binned)

(c) Network *'Union'*

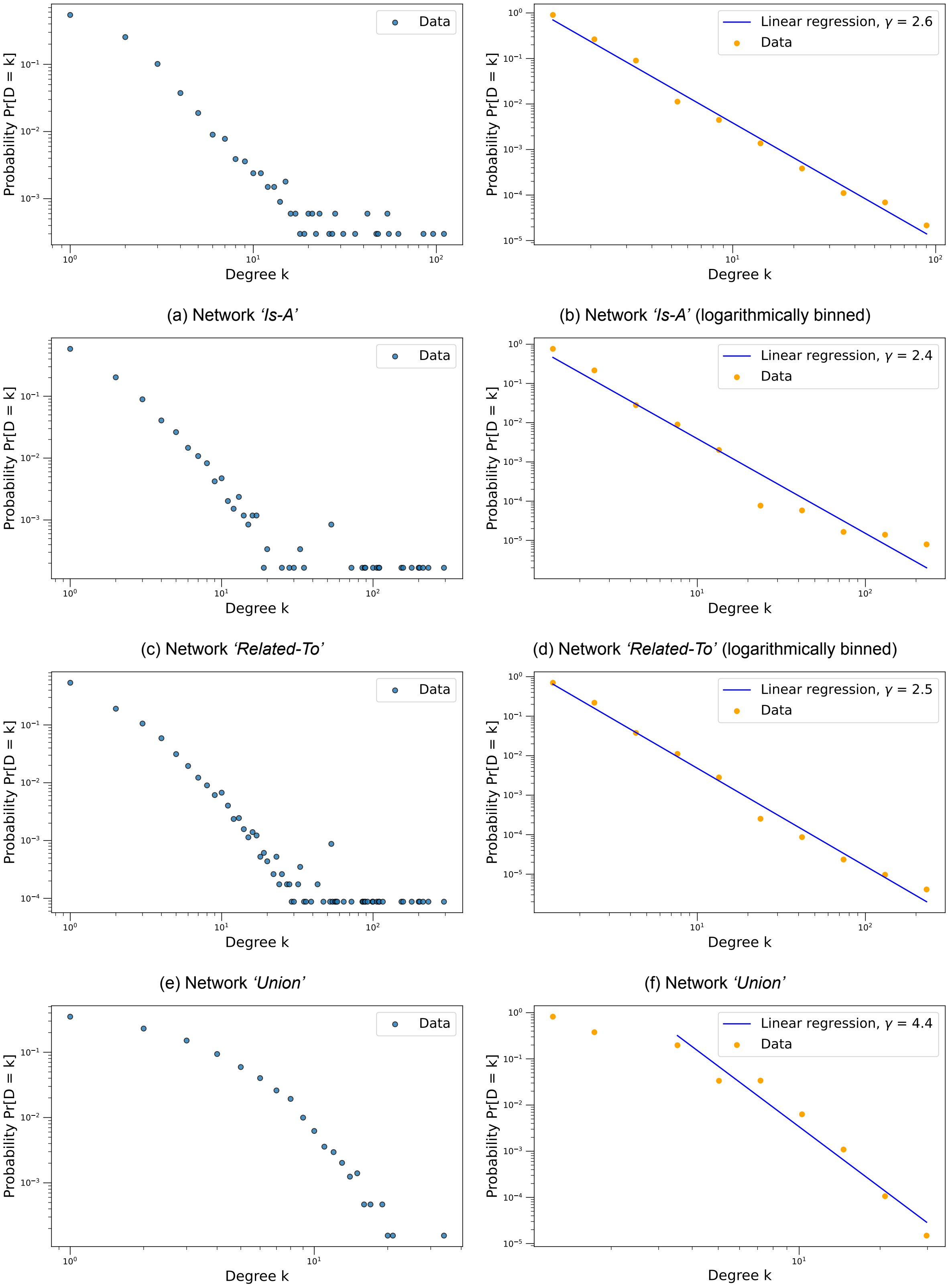
(d) Network *'Union'*

(e) Network *'Synonym'*

(f) Network *'Synonym'*

Figure D.5: Degree distributions of three Russian semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.6. Portuguese



(a) Network *'Is-A'*

(b) Network *'Is-A'* (logarithmically binned)

(c) Network *'Related-To'*

(d) Network *'Related-To'* (logarithmically binned)

(e) Network *'Union'*

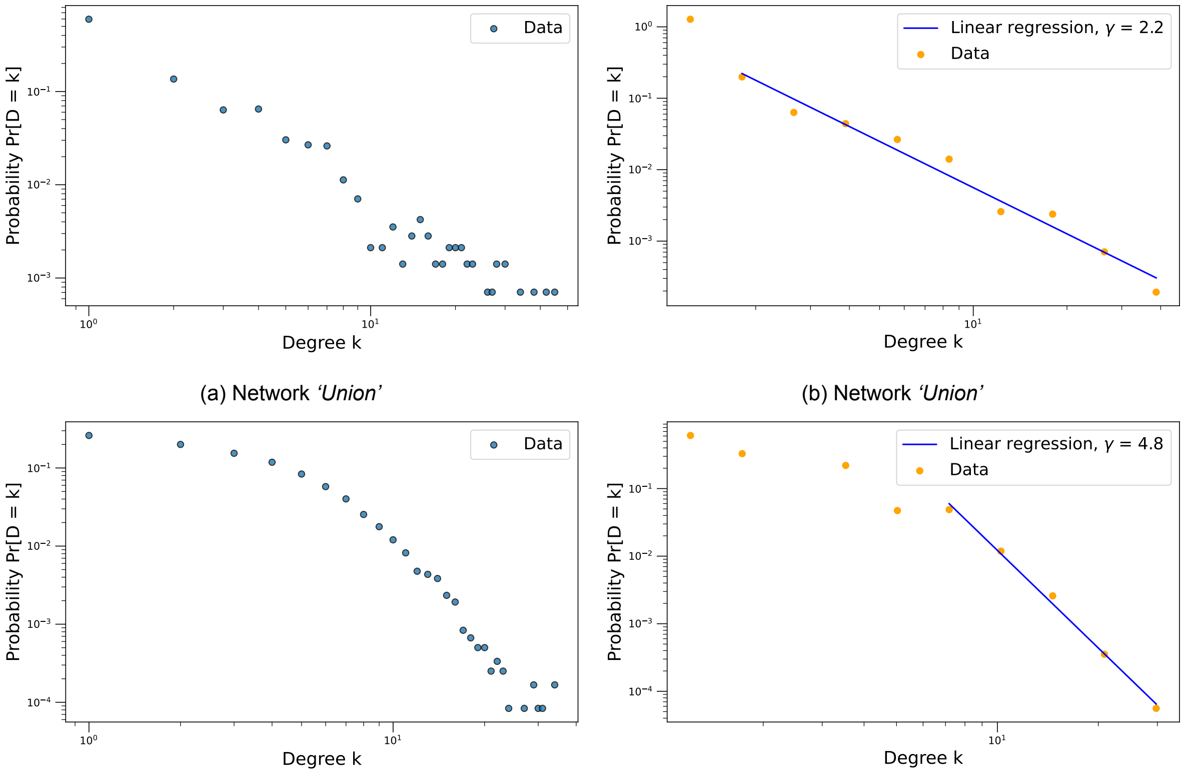
(f) Network *'Union'*

(g) Network *'Synonym'*

(h) Network *'Synonym'*

Figure D.6: Degree distributions of four Portuguese semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.7. Dutch



(a) Network *'Union'*

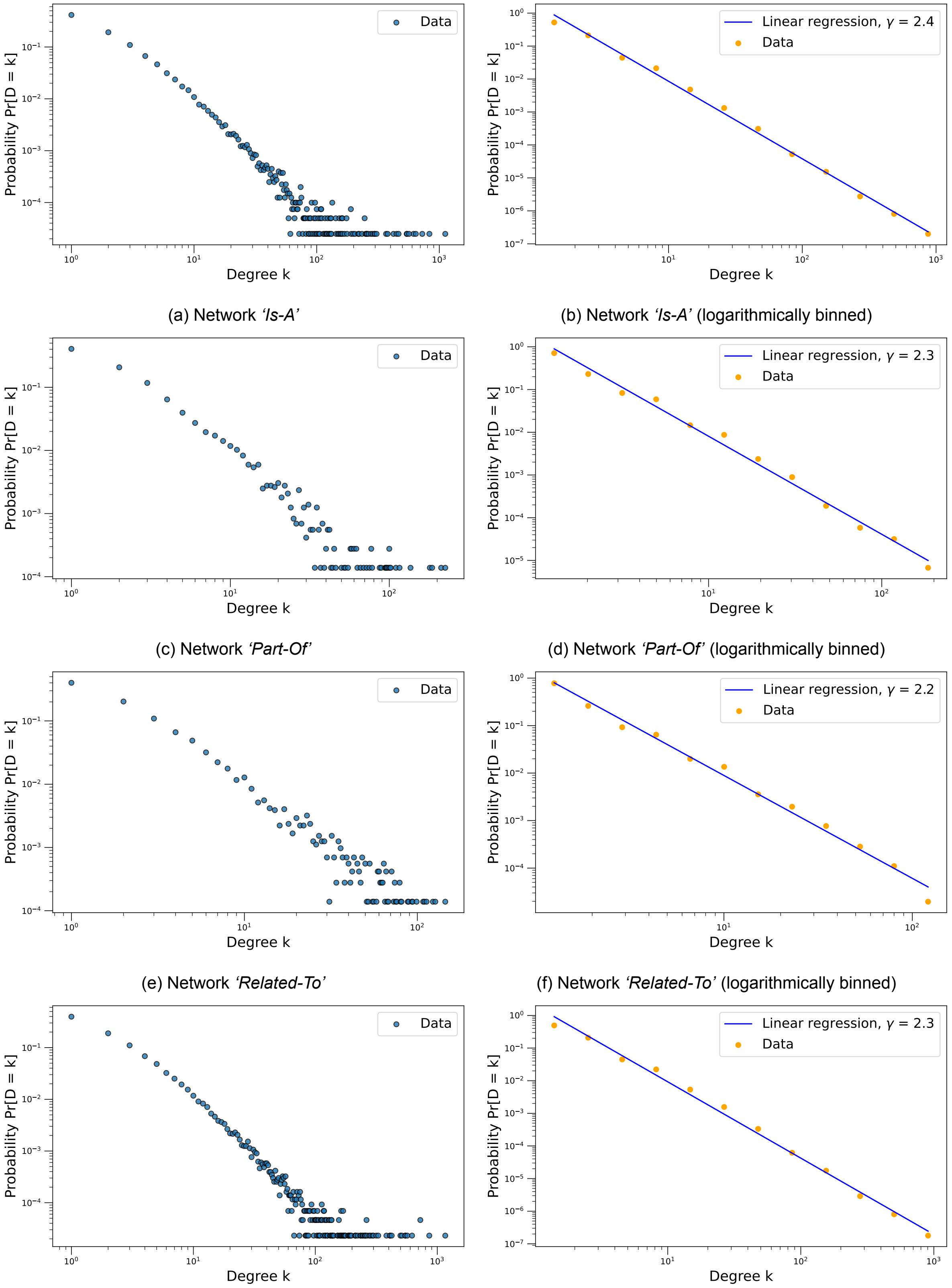
(b) Network *'Union'*

(c) Network *'Synonym'*

(d) Network *'Synonym'*

Figure D.7: Degree distributions of two Dutch semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.8. Japanese



(a) Network *'Is-A'*

(b) Network *'Is-A'* (logarithmically binned)

(c) Network *'Part-Of'*

(d) Network *'Part-Of'* (logarithmically binned)

(e) Network *'Related-To'*

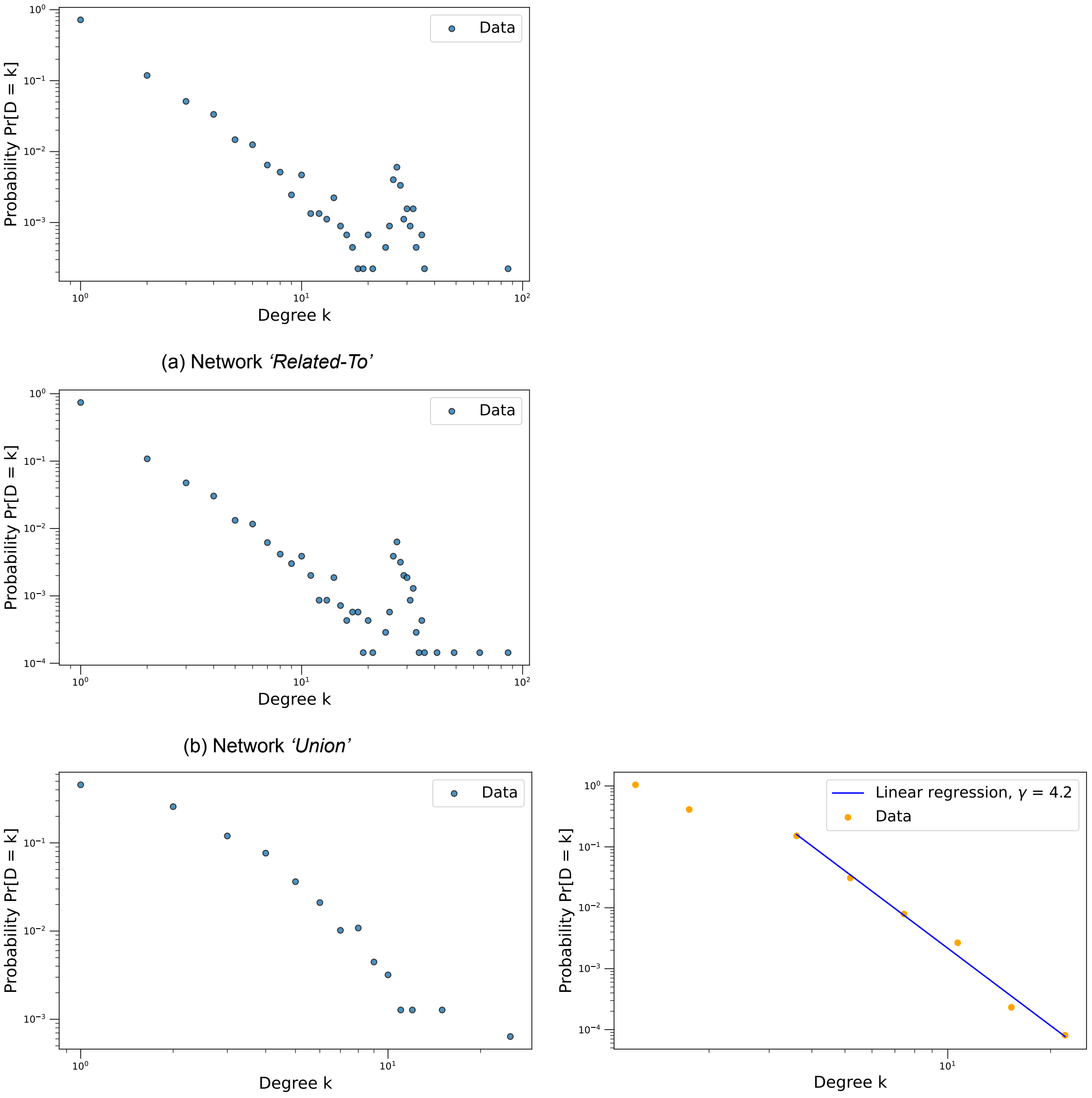
(f) Network *'Related-To'* (logarithmically binned)

(g) Network *'Union'*

(h) Network *'Union'*

Figure D.8: Degree distributions of four Japanese semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.9. Finnish



(a) Network *'Related-To'*

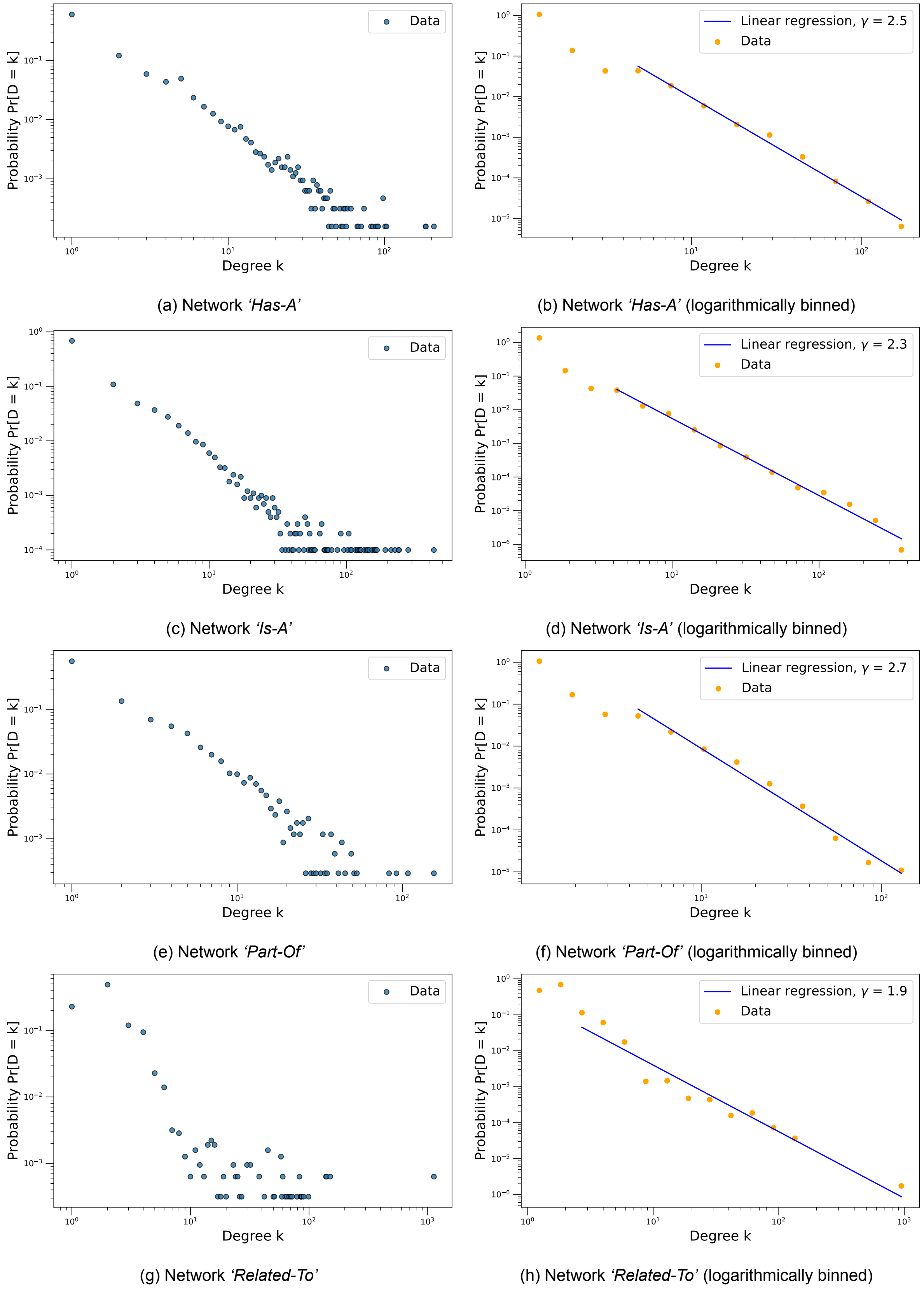
(b) Network *'Union'*

(c) Network *'Synonym'*

(d) Network *'Synonym'*

Figure D.9: Degree distributions of three Finnish semantic networks and power-law exponents estimation over logarithmically binned degree distribution.

D.10. Chinese



(a) Network *'Has-A'*

(b) Network *'Has-A'* (logarithmically binned)

(c) Network *'Is-A'*

(d) Network *'Is-A'* (logarithmically binned)

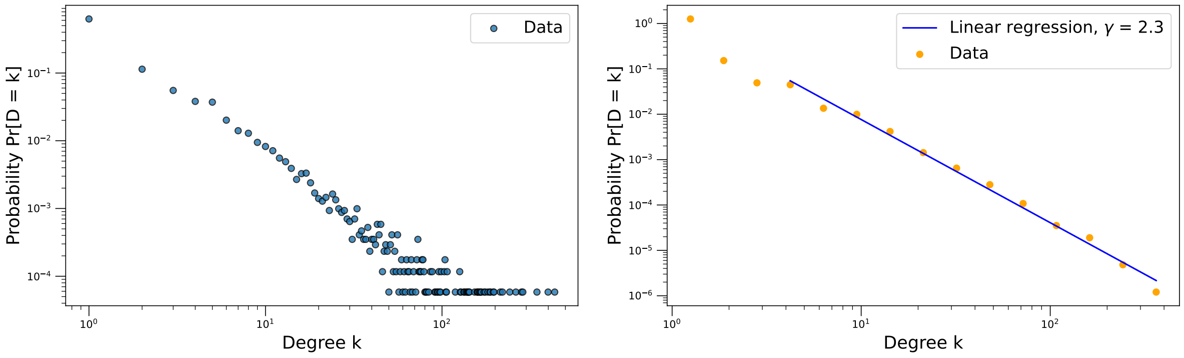
(e) Network *'Part-Of'*

(f) Network *'Part-Of'* (logarithmically binned)

(g) Network *'Related-To'*

(h) Network *'Related-To'* (logarithmically binned)

Figure D.10: Degree distributions of five Chinese semantic networks and power-law exponents estimation over logarithmically binned degree distribution.



(i) Network *'Union'*

(j) Network *'Union'*

Figure D.10: Degree distributions of five Chinese semantic networks and power-law exponents estimation over logarithmically binned degree distribution (cont.).

E

Appendix

This appendix presents the configuration model used to calibrate structural coefficients of semantic networks. The details of the calibration process are provided as well.

E.1. Undirected Binary Configuration Model

In this thesis, we utilize Undirected Binary Configuration Model (UBCM) [87] to calibrate the structural coefficients. The UBCM generates a maximum entropy probability distribution over a network with the constraints of an expected degree sequence. It is suitable for undirected and unweighted networks. The resulting maximum entropy distributions are maximally unbiased with respect to any other property [88].

E.2. Details of Calibration Process

Networks that have fewer than 100 nodes are skipped, because there is a high chance that there exist no triangles or quadrangles in a sampled network. As a result, the structural coefficient $x(G_i) = 0$. When $x(G_i) = 0$, Eq. 5.9 is undefined.

Since the runtime of the algorithm depends on the size of a network and the choice of the number of randomized networks R, we skip the two largest networks French 'Related-To' and 'Union' with $N > 1,200,000$ due to limited time. Because the calibrated values are arithmetic averages taken over independent samples, intuitively, any R between 100 and 1000 should give a relatively good estimation of the null distribution of a structural coefficient. We also validate that the calibrated values obtained using $R = 100$ and $R = 500$ differ mostly at the hundredths. Therefore, we use $R = 500$ for most networks. Only for the two large networks, English 'Related-To' and 'Union', we use $R = 100$ to avoid long computation time.