Measuring Diversity in Heterogeneous Information Networks

Pedro Ramaciotti Morales

Sciences Po, médialab, Paris, France

Robin Lamarche-Perrin

CNRS, ISC-PIF, LIP6, Paris, France

Raphaël Fournier-S'niehotta

CEDRIC, CNAM, Paris, France

Rémy Poulain, Lionel Tabourier

Sorbonne Université, CNRS, LIP6, Paris, France

Fabien Tarissan

Université Paris-Saclay, CNRS, ISP, ENS Paris-Saclay, Cachan, France

Abstract

Diversity is a concept relevant to numerous domains of research as diverse as ecology, information theory, and economics, to cite a few. It is a notion that is continuously gaining attention in the information retrieval, network analysis, and artificial neural networks communities. While the use of diversity measures in network-structured data finds a growing number of applications, no clear and comprehensive description is available for the different ways in which diversities can be measured in relational data described in these structures. In this article, we develop a formal framework for the application of a large family of diversity measures to heterogeneous information networks (HINs), a flexible, widely used, network data formalism. This allows for an extension of the application of diversity measures, from systems of classifications and apportionments, to more complex relations that can be better modeled by networks. In doing so, we do not only provide an effective organization of multiple practices from different domains, but we also unearth new observables in systems modeled by heterogeneous information networks. The pertinence of the approach is illustrated by the development of different applications related to various domains concerned by both diversity and networks. In particular, we illustrate the usefulness of these new proposed observables in the domains of recommender systems and new media studies among other fields.

Key words: diversity measures, heterogeneous information networks, random walks on graphs, recommender systems, social network analysis.

Contents

1 Introduction **3**

2 The Concept of Diversity **4**
2.1 Items, types, and classifications 4
2.2 The diversity of diversity measures 5
2.3 A theory of diversity measures 6
2.3.1 Properties of diversity measures 7
2.3.2 Self-weighted quasilinear means 8
2.3.3 True diversities 8
2.4 Relative true diversities 10
2.5 Joint distributions, additivity, and Shannon entropy 11

3 Random Walks in Heterogeneous Information Networks **12**
3.1 Preliminary notations 12
3.2 Heterogeneous information networks 13
3.3 Meta paths and constrained random walks 14

4 Network Diversity Measures **16**
4.1 Collective and individual diversity 16
4.2 Backward diversity 18
4.3 Relative diversity 21
4.4 Projected diversity 21
4.5 Summary of network diversities measures 23

5 Applications **23**
5.1 A simple example 23
5.2 Recommender Systems 24
5.3 Social media studies, echo chambers, and filter bubbles 26
5.4 Other examples 28

6 Conclusions **31**

1. Introduction

Diversity is a concept of importance in several different domains of research, such as ecology [1], economics, public policy [2], information theory [3, 4], new media studies [5, 6], and complex system [7, 8] among many others. Across the full range of domains where it is used, diversity refers to some combination of three properties of systems, identified as *variety* (the number of types of entities), *balance* (the distribution of entities into types), and *disparity* (the difference between types of entities) [9]. Diversity measures are quantitative indices for these properties. Prominent examples are the Shannon entropy in information theory [10], the Gini Index in economics [11], or the Herfindahl-Hirschmann Index [12] in competition law. Examples of the application of these indices can be found in the measurement of biodiversity in ecology [13], of industrial concentration in economics [14, 15], or in the measurement of online social phenomena such as filter bubbles and echo chambers [5]. In the context of digital platforms and online medias, the notion of diversity has recently become central as well. The fact that digital platforms increasingly resort to the use of algorithmic recommendations to drive users choices has led the scientific community to analyze the impact of the recommendations to the users. If one can argue that this recent development provides the users with useful information, the phenomenon also feeds into fears about unpredictable outcomes over the long term, the most debated one begin the emergence of the so-called *filter bubbles* [16, 17, 18]. In this context, while the need to measure and audit recommendation systems is commonly agreed upon [19, 20], there is no consensus on how to properly measure the impact of the recommendations on the users. However, many studies have highlighted the need to explore diversity or serendipity (the fortunate and unexpected discovery of unexpected items) in the way information is exposed to the users [21, 22, 23].

Diversity measures can be computed over different types of data and in a multitude of contexts. Access to data on traces or representations of different real phenomena has allowed for an explosive extension of the reach of quantitative studies in many disciplines. One particular type of data over which diversity measures can be computed is network-structured data, best represented using graph formalisms. Recently, formalisms such as the *heterogeneous information networks* (HINs) [24, 25] have been used successfully to provide ontologies for unstructured data, especially in the contexts of information retrieval [26] and recommender systems [27], as well as in the artificial intelligence and representation learning communities [28, 29, 30].

Much of the success of these representations and their precursors – such as the multi-layer graphs [31, 32] – is due to the way in which semantic relations can be mapped to sets of paths between groups of entities according to their type (called *meta paths*), and to the the way in which they can be used in algorithms. One prominent way of exploiting meta paths is by constraining random walks inside them. This has been extensively used in the computation of similarity [33, 34, 35] or for recommendation purposes [27, 36, 37]. While the application of diversity measurements to graph structures is not new [38, 39], its is gaining widespread use in different communities [40], and in particular in the information retrieval and recommender systems communities [41]. A few works have hinted at the application of entropy (one prominent diversity measure) to distributions computable from meta path structures in heterogeneous information networks. This has been done to provide diverse recommendations [42]. In similarity search, entropy has also been used to measure information gain in the selection of different meta paths [43, 44]. However, no clear and comprehensive description is currently available for the different ways in which diversity measures can be computed form data described with network structures. On the one hand, there is no clear organization or description of the different distributions computable from meta paths over which to measure the diversity. On the other hand, this practice remains largely limited to applications in information theory, used to decide inclusion of information in similarity computations. Several communities interested in both network representation models and diversity measures have limited – or no – examples of application at their disposal, let alone any theory or a framework on which to develop applications.

In this article we develop *network diversity measures*: a comprehensive theory of diversity and a formal framework for its application to network-structured data. This framework relies on modeling data with heterogeneous information networks using multigraphs for generality. Doing so, we collect and unify a wide range of results in quantitative diversity measures across different disciplines covering most practical uses. In developing this formal framework, we provide a description of several practices existing in the scientific literature. In addition, we also point to new information to be extracted by measuring the diversity of previously unconsidered observables in network-structured data. One of the main applications of the network diversity measures is the extension of diversity measures, from relatively simple systems of classification and apportionment (e.g. species in ecosystems for ecology, units produced by firms in economy) to more complex data best modeled by network structures. The relevance and usefulness of these new network diversity measurements are illustrated by the development of practical examples from different domains of research, including recommender systems, social media and platforms, and ecology, among others.

The main contributions of this article are:

- A new organization of an axiomatic theory of diversity measures encompassing most uses across several disciplines.
- A formalization of concepts emerging in graph theory (specially in applications in recommender systems, information retrieval, and representation learning communities), in particular that of the meta path and different observables computable from it.
- The proposition of several *network diversity measures*, resulting from the of application diversity measures to distribution probabilities computable in the heterogeneous information network formalism.
- The application of these network diversity measures to previously existing quantitative observables in different research domains, and the development of new applications through examples.

In Section 2, we provide a framework to present and organize the diversity measures of the literature. This framework has the advantage of covering a large part of the existing concepts around diversity, and that of formalizing the algebraic properties that they obey. Then, in Section 3, we define random walks in the context of heterogeneous information networks. In particular, we formalize the concept of meta path. Constrained random walks along particular meta paths will play a central role in the rest of the article when computing diversity in systems represented by networks. Indeed, in Section 4, we combine the diversity measures described within the framework to different observables computed from constrained random walks in order to derive families of interpretable network diversity measures. Finally, Section 5 illustrates the relevance of these measures by using them in applications from various fields concerned by the concept of diversity.

2. The Concept of Diversity

In a general sense, diversity refers to some properties of a system that contains items that are classified into types. These properties are related to the number of types used, the way in which items are classified into types, and how different are types between them. This simple model of items classified into types accounts for the usage of diversity in many domains of research. Prominent examples are units of wealth or revenue classified as belonging to different persons (in economics), number of individuals classified into different species (in ecology), or produced units of a commodity classified by firms (in competition law).

2.1. Items, types, and classifications

Let us consider a system made of a set I of items, a set T of types, and a membership relation $\tau \subseteq I \times T$ indicating items that are classified into types: item $i \in I$ is classified as being of type $t \in T$ if and only if $(i, t) \in \tau$. A diversity measure is a function $D : I \times T \to \mathbb{R}^+$ that maps a system to a diversity value d, i.e., $D : \tau \mapsto d \in \mathbb{R}^+$. In this article we consider diversity measures for systems with a given set of items, a given set of types, and a given classification of item into types.

We do not consider the problem of identification, i.e., what should be considered as item in a universe of possible elements, and what types should be considered for a classification. This is an important question, however it deals rather with the meaning of the elements of the system in terms of semantic, which is out of the scope of this work.

We define the *abundance* of type $t \in T$ as the number of items of that type: $a_\tau(t) = |\{i \in I : (i, t) \in \tau\}|$, and the *proportional abundance* as $p_\tau(t) = \frac{a_\tau(t)}{|\tau|}$. Using these definitions, we further narrow our consideration of diversity measures to functions that map proportional abundances resulting from a classification to non-negative real values: $D(\tau) = D(p_\tau(t_1), \ldots, p_\tau(t_k))$ with $k = |T|$ being the number of types. Hence, a diversity measure D is an application from Δ^* to $\mathbb{R}^+$, where $\Delta^* = \cup_{k\geq 0}\Delta^k$ is the union of all standard k-simplices, that is the sets of probability distributions on discrete spaces of size $k + 1$:

$$\Delta^k = \left\{(p_1, \ldots, p_{k+1}) \in \mathbb{R}^{k+1} : \forall i \leq k,\ p_i \in [0, 1] \text{ and } \textstyle\sum_{i\leq k+1} p_i = 1\right\}.$$

2.2. *The diversity of diversity measures*

As stated in the previous subsection, the term *diversity* is used to designate various properties of dissimilarity in a range of domains, such as ecology [13, 45, 46, 47, 48, 49, 1], life sciences [50], economics [2, 12], public policy [51, 52, 53, 54], information theory [3, 4], internet & media studies [5, 6], physics [55, 56], social sciences [57], complexity sciences [58, 59, 7, 8], and opinion dynamics [60]. Across the domains in which it is used, this term refers to different properties of systems of items classified into types. Accordingly, diversity measures are functions assigning to each system a diversity value, intended to be a quantitative measurement of these different properties.

The properties intended to be referred by the term *diversity* across the full range of sciences mentioned are some combination of three properties, identified as *variety*, *balance*, and *disparity* [9]:

- *variety* is the number of types into which the items of a system can be classified;
- *balance* is a measure of the extent to which the pattern of the proportional abundance resulting from a classification of items into types is evenly distributed (i.e., balanced);
- and *disparity* is the degree to which different types can be differentiated according to a metric for the set of types T.

The reader is referred to [61] for an extended discussion on these properties.

We illustrate the concept of *diversity* through classic examples of diversity measures present in works from different fields. For this purpose, we consider a proportional abundance distribution $p = (p_1, p_2, \ldots, p_{|T|})$ resulting from the classification of items I into types T.

Richness [62, 63] is a common diversity measure, only related to the property of *variety*. Often used in ecology, it simply measures the number of types *effectively* used to classify items. If a bookcase contains novels, comics, and travel books, its richness is equal to 3, whatever their proportions may be.

$$R(p) = |\{i \in \{1, 2, ..., |T|\} : p_i > 0\}|.$$

Richness only counts the types effectively used in a classification. If one considers a classification with 20 possible types to examine two bookcases, the first having books of 3 types and the second of 4 types, the second one will be more diverse under this measure. The property captured by this measure coincides with the property identified as *variety*. Richness can serve as a basis for other measures. The ratio between richness and the number of classified items is one example [45, Section 9].

Shannon entropy [10, 64], here denoted by H and related to the *variety* and *balance* properties, is a cross-disciplinary diversity measure, but most regularly found in the field of information theory. It quantifies the uncertainty in predicting the type of an item taken at random. If one knows the proportional abundance of types of books in a bookcase, and if one takes books from it at random, Shannon entropy is the average number of type-checks one would have to make per book in order to determine its type.

$$H(p) = -\sum_{i=1}^{|T|} p_i \log_2 p_i.$$

Many classical diversity measures are functions of the properties identified as *variety* and *balance*. Shannon entropy, introduced in the context of channel capacity in telecommunications, is clearly affected by the proportional abundances – and thus by the *balance* – of a system, but also by its *variety*: according to Shannon entropy, a bookcase with books that are uniformly distributed among 5 types is more diverse than a bookcase with books that are uniformly distributed among 4 types. But applying normalization, one can restrain the measurement to the property of *balance*. The diversity measure known as **Shannon Evenness** [65], for example, consisting on the ratio between the measured entropy and the maximal possible entropy for the same number of effective types, only accounts for the property of *balance*.

Herfindahl-Hirschman Index [12], here denoted as HHI, is mainly used in competition law or antitrust regulation in economy. It is intended to measure the degree of concentration of items into types. If one knows the proportional abundance of types of books in a bookcase, and if one takes 2 books from it at random, the Herfindahl-Hirschman Index is the probability of them being of the same type.

$$\mathrm{HHI}(p) = \sum_{i=1}^{|T|} p_i^2.$$

Also related to the *variety* and *balance* properties, this index (also known as the **Simpson Index** [66]), was first introduced by Hirschman [67] and later by Herfindahl [68] in the study of the concentration of industrial production.

A related diversity measure, the **Gini-Simpson Index** [11] (also called **Gibbs-Martin Index** in sociology and psychology [69] and **Population Heterozygosity** in genetics [70]), is another prominent example of a measure accounting for *variety* and *balance*. Also known as the probability of interspecific encounter in ecology [71], it is the probability of the complementary event associated with the Herfindahl-Hirschman Index, i.e., the probability of taking at random two items with different types.

This is not to be confused with the **Gini Coefficient** [72], commonly used in economics, which is a *balance*-only diversity measure that can be interpreted as a measure of inequality when items are units of wealth distributed into types. One of the formulations of the Gini Coefficient is given by

$$\mathrm{Gini}(p) = \frac{1}{2|T|} \sum_{i=1}^{|T|} \sum_{j=1}^{|T|} |p_i - p_j|.$$

Other diversity measures address only the property of *balance*. The **Berger-Parker Index** [73], here denoted as BPI, is another prominent example. Also common in ecology, it measures the proportional abundance of the most abundant type. If 90% of the books in a bookcase are comics, its Berger-Parker Index will be 0.9, and for whatever classification given to the remaining 10% of the books.

$$\mathrm{BPI}(p) = \max_{i \in \{1,2,\ldots,|T|\}} p_i.$$

It is easy to see that only the *balance* property affects this diversity measure. If the books of a bookcase can be classified as 90-10% into two types according to one classification, but as 90-5-5% into three types according to another one, both classifications still have the same diversity according to this measure.

Another group of existing diversity measures addresses the *disparity* property. In its most general form, a pure-*disparity* diversity measure is a function of the pairwise distance between types of T in some disparity space [74, 75]. *Disparity* is the property underlying a considerable amount of cases of the use of the notion of diversity. Examples of this can be found in fields such as paleontology [76], economics [77], and biology [78]. Furthermore, diversity measures accounting for *disparity* as well as *variety* and *balance* exist [79].

While the measurement of *disparity* relies on the existence of topological or metrical structures for the set of types T, that of *variety* and *balance* rely solely on the establishment of identification and classification in a system of items and types, which is the setting of many studies and applications. As indicated in the previous subsection, in the rest of this article we will focus on diversity measures for these kinds of settings, thus leaving aside the *disparity*-related diversity measures.

2.3. A theory of diversity measures

In Section 2.1, we first limited the scope of diversity measures to that of functions mapping systems with given items, types, and a classification, to non-negative real numbers. Then we limited the scope further, to functions mapping probability distributions to non-negative real numbers. In this section, we further reduce the scope of diversity measures by prescribing axioms reflecting on desired properties such measures should have.

In the domain of information theory, there are several possible axiomatic theories that give rise to entropies and diversity measures (cf. [80, 81, 82, 83]). Drawing from existing axiomatizations, we propose an organization of axioms suited for the purposes of this article.

We first introduce four axioms that encode properties which are necessary for a diversity measure, i.e. *symmetry*, *expansibility*, *transferability*, and *normalization*. Then, we present a family of functions that satisfies these properties. By imposing the additional property known as *replicability*, in the form of an axiom, the resulting measures of the theory correspond to the family of functions known as *true diversities*. One among them, closely related to the Shannon entropy, has additional properties of interest for the measurement of diversity in networks.

2.3.1. Properties of diversity measures

Let us consider a diversity measure as a function $D : \Delta^{k-1} \to \mathbb{R}^+$, a probability distribution $p = (p_1, ..., p_k) \in \Delta^{k-1}$, and some properties of interest in the form of axioms for a theory of diversity.

A first property, called *symmetry* (or *anonymity*), is said to be satisfied by a diversity measure if it is invariable to permutation of items between types. For instance, a bookcase with 25% of comics and 75% of novels has the same diversity than a bookcase with 75% of comics and 25% of novels under a symmetric diversity measure. It means that symmetric diversity measures are blind to the nature of the types.

Axiom 1 (Symmetry) *For any permutation σ on the set $\{1, 2, ..., k\}$, a diversity measure D is symmetric if*

$$D(p_1, p_2, ..., p_k) = D(p_{\sigma(1)}, p_{\sigma(2)}, ..., p_{\sigma(k)}).$$

We also demand of a diversity measure to be *expansible*, or *invariant to non-effective types*, i.e., types with no items. Adding a type with no items does not impact the diversity: in the same bookcase, considering the type dictionaries, with 0% of books, does not change its diversity.

Axiom 2 (Expansibility) *A diversity measure D is expansible if*

$$D(\underbrace{p_1, p_2, ..., p_k}_{k \text{ entries}}) = D(\underbrace{p_1, p_2, ..., p_k, 0}_{k+1 \text{ entries}}).$$

For a diversity measure to be a measure of balance, it needs to satisfy the *transfer* –or *Pigou-Dalton*– *principle* [84]: if a bookcase has more novels than comics, replacing some novels with new comics should increase the diversity (if the new number of comics does not surpass the new number of novels).

Axiom 3 (Transfer Principle) *A diversity measure D satisfies the transfer principle if for all i, j in $\{1, ..., k\}$, if $p_i > p_j$, then*

$$\forall \epsilon \leq \frac{p_i + p_j}{2}, \quad D(\underbrace{\ldots, p_i - \epsilon, \ldots, p_j + \epsilon, \ldots}_{k \text{ entries}}) \geq D(\underbrace{\ldots, p_i, \ldots, p_j, \ldots}_{k \text{ entries}}).$$

It is easy to verify that axioms 1, 2, and 3 imply the following *merging* property.

Theorem 1 (Merging) *A diversity measure D that satisfies axioms 1, 2 & 3 is such that*

$$D(\underbrace{\ldots, p_i, p_{i+1}, \ldots}_{k \text{ entries}}) \quad \geq \quad D(\underbrace{\ldots, p_i + p_{i+1}, \ldots}_{k-1 \text{ entries}}).$$

These first three axioms also imply that diversity measures of the theory are bounded.

Theorem 2 (Bounds for diversities measures) *A diversity measure D –that satisfies axioms 1, 2 & 3– is such that*

$$D(\underbrace{1/k, 1/k, \ldots, 1/k}_{k \text{ entries}}) \quad \geq \quad D(p_1, p_2, \ldots, p_k) \quad \geq \quad D(\underbrace{1, 0, \ldots}_{k \text{ entries}}).$$

In order for diversity measures to have a scale for their measurement, we need to impose values to the minimal and maximal diversity [15]. We establish this as a property, called the *normalization* principle. Normalization means that if all types of books are equally abundant in a bookcase, the diversity is equal to the number of effective types.

Axiom 4 (Normalization) *A diversity measure D satisfies the normalization principle if*

$$D(\underbrace{1/k, ..., 1/k}_{k \text{ entries}}) = k.$$

It is easy to see that the diversity values of the theory are bounded as a consequence of the normalization axiom.

Theorem 3 (Bounds for diversity values) *A diversity measure D –that satisfies axioms 1, 2, 3 & 4– is such that, for all $p \in \Delta^{k-1}$, we have $k \geq D(p) \geq 1$.*

2.3.2. *Self-weighted quasilinear means*

One of the advantages of restricting the scope of diversity measures to functions of distributions $p \in \Delta^*$, is that they can be used in conjunction with probability computations, as it will be shown in Section 4. These measures considered thus far also belong to the more general class of aggregation functions. The most general form of aggregation function that is compatible with the axioms of probability [85] is the family of quasilinear means (derived by Kolmogorov [86] and Nagumo [87]). Quasilinear means are central to quantification of information in information theory [3], and are of the form

$$D(p) = \phi^{-1}\left(\sum_{i=1}^{k} w_i \phi(p_i)\right),$$

for weights w_i such that $\forall i \in \{1, \ldots, k\}, (0 \leq w_i \leq 1)$ with $\sum_{i=1}^{k} w_i = 1$, and for ϕ a strictly monotonic increasing continuous function.

A sub-family of quasilinear means, the so-called self-weighted quasilinear means, has additional properties that will be of interest in what follows. They are formally defined as follows.

Definition 1 (Self-weighted quasilinear means [88]) *A function $D : \Delta^* \to \mathbb{R}^+$ is a self-weighted quasilinear mean if it can be written as*

$$D(p) = \phi^{-1}\left(\sum_{i=1}^{k} p_i \phi(p_i)\right),$$

with ϕ a strictly monotonic increasing continuous function.

Further restrictions of the considered diversity measures, described by the following theorem, result in a family of functions that satisfy both the desired properties stated as axioms, and the axioms of probability.

Theorem 4 (Concave self-weighted quasilinear means are diversity measures of the theory [15, 14]) *Self-weighted quasilinear means D, such that $h(t) = t\,\phi(t)$ is concave (with function ϕ from Definition 1), satisfy Axioms 1, 2, 3 & 4.*

Self-weighted quasilinear means is a subset of the functions defined by Axioms 1, 2, 3 & 4. Indeed, there are diversity measures that satisfy these axioms but are not self-weighted quasilinear means (e.g. the Hall-Tideman Index [89]).

2.3.3. *True diversities*

An additional property, the *replication principle*, captures a characteristic of some diversity measures according to which, if types are replicated $m \geq 1$ times, the diversity value is multiplied by m [15]. Let us suppose, for example, that 25% of a bookcase is made of comics and 75% is made of novels. Let us also suppose that we add new items from a different bookcase, of which 25% are dictionaries and 75% are photo albums. The diversity of the new –replicated– bookcase with four types of books is double that of that of the original bookcase. The replication principle is needed to avoid otherwise paradoxical results in many applications [90].

Axiom 5 (Replication) *A diversity measure D satisfies the replication principle if it is such that*

$$D\left(\underbrace{\frac{p_1}{m}, \frac{p_2}{m}, \ldots, \frac{p_k}{m}}_{1^{st}\ copy}, \underbrace{\frac{p_1}{m}, \frac{p_2}{m}, \ldots, \frac{p_k}{m}}_{2^{nd}\ copy}, \ldots, \underbrace{\frac{p_1}{m}, \frac{p_2}{m}, \ldots, \frac{p_k}{m}}_{m^{th}\ copy}\right) = m\, D(p_1, \ldots, p_k).$$

The addition of the replication principle to the theory of diversity uniquely defines a sub-family within that of the concave self-weighted quasilinear means, called *true diversities*.

Definition 2 (True diversity of order α) *The true diversity or order α, or α-order true diversity, denoted D_α, is an application $D_\alpha : \Delta^* \to \mathbb{R}^+$, such that, given $p = (p_1, \ldots, p_k) \in \Delta^*$ and $\alpha \in \mathbb{R}^+$, is defined as*

$$D_\alpha(p) = \left(\sum_{i=1}^{k} p_i^\alpha\right)^{\frac{1}{1-\alpha}} \quad \text{if } \alpha \neq 1, \quad \text{and} \quad D_1(p) = \left(\prod_{i=1}^{k} p_i^{p_i}\right)^{-1}, \quad \text{with } {p_i}^{p_i} := 1 \text{ if } p_i = 0.$$

Note that variants of true diversities (first introduced as the Hill Number [13] and named *true diversity* in [91]) exist in different domains. The Hannah-Kay concentration index of order α [92] is the reciprocal of D_α. In information theory, the Rényi Entropy [4] of order α, denoted by H_α , is the natural logarithm of D_α: $H_\alpha(p) = \ln D_\alpha(p)$.

Theorem 5 (Diversity measures the theory that satisfy the replication principle are true diversities [15]) *Suppose that a diversity measure D can be represented as a self-weighted quasilinear mean. Then D is a true diversity for some order α if and only if D satisfies the replication principle of Axiom 5.*

It must be remembered that, despite its name, true diversity is a index –that respects given axioms– for the properties included in diversity, and not diversity itself nor any other phenomenological entity [93].

True diversities are related to several of the diversities used in different domains and identified in Section 2.2. The *richness* of a distribution p can be computed as the limit of $D_\alpha(p)$ when $\alpha \to 0^+$, observing that $p_i^\alpha \to 1$ if $p_i > 0$, thus resulting in the count of effective types. We thus identify richness with the 0-order true diversity, calling it **Richness diversity**. $D_1(p)$, the 1-order true diversity (or **Shannon diversity**), also called *perplexity* [94], is related to the Shannon entropy $H(p)$ of p by exponentiation: $D_1(p) = 2^{H(p)}$ when the entropy is computed in base 2. $D_2(p)$, the 2-order true diversity (or **Herfindhal diversity**), is the reciprocal of the Herfindhal-Hirschman Index: $D_2(p) = 1/\mathrm{HHI}(p)$. The Berger-Parker Index can also be identified with the result of a limit process. By observing that $D_\alpha \xrightarrow{\alpha\to\infty} 1/\max\{p_1, \ldots, p_k\}$ (Section 5.4 of [3]) we can define

$$D_\infty(p) := \frac{1}{\max\{p_1, \ldots, p_k\}},$$

and thus conclude that $D_\infty(p) = 1/\mathrm{BPI}(p)$ (here called **Berger diversity**). These relations are summarized in Table 1. In the previous relations, the fact that Herfindhal-Hirschman Index and the Berger-Parker Index are reciprocal to true diversities underlines that they are intended to measure concentration.

Table 1: Summary of true diversities of order 0, 1, 2, and ∞, and their relation to classic diversity measures.

Order (α)	Name	True diversity	Expression	Relation to other diversity measures
0	Richness diversity	$D_0(p)$	$\lvert\{i \in \{1, \ldots, k\} : p_i > 0 \in\}\rvert$	Same as richness [62, 63].
1	Shannon diversity	$D_1(p)$	$\left(\prod_{\substack{i=1 \\ p_i\neq 0}}^{k} p_i^{p_i}\right)^{-1}$	Exponential of Shannon entropy [10, 64]: $H(p) = \log_2(D_1(p))$, with H in base 2.
2	Herfindahl diversity	$D_2(p)$	$\left(\sum_{i=1}^{k} p_i^2\right)^{-1}$	Reciprocal of the Herfindahl-Hirschman Index [12]: $\mathrm{HHI}(p) = 1/D_2(p)$.
∞	Berger diversity	$D_\infty(p)$	$\left(\max_{i\in\{1,\ldots,k\}} \{p_i\}\right)^{-1}$	Reciprocal of Berger-Parker Index [73]: $\mathrm{BPI}(p) = 1/D_\infty(p)$.

Let us illustrate some of these properties in Figure 1. By virtue of the axioms of the theory, all true diversities have equal values for uniform distributions with the same number of effective (non-empty) types. In this case, diversity is the number of effective types (horizontal lines in Figure 1). However, when the distribution into types is not uniform, these measures behave differently (decreasing curves in Figure 1). In this case, parameter α expresses the way non-uniformity, or *balance* is taken into account. If α is low, inequalities in the distribution will only have a weak impact on the diversity value, and in the extreme case where $\alpha = 0$ (i.e. for richness), inequalities in proportional abundances are not at all taken into account. Conversely, if α is high, inequalities in the distribution will have a strong impact on the diversity value, and in the extreme case where $\alpha \to \infty$ (i.e. for the Berger diversity), only the highest abundance is taken into account. The red and blue curves in Figure 1 illustrate how the parameter α can modulate the relative importance given to the properties of *variety* and *balance* (cf. Section 2.2): a distribution with 6 types can be less diverse than one with 4 types if it is sufficiently unbalanced for a given value of α. True diversities allow us to have a continuum of measures which give a different weight to *variety* and *balance* of a distribution: $\alpha \to 0$ means that diversity takes only *variety* into account, while $\alpha \to \infty$ means that diversity takes only *balance* into account.

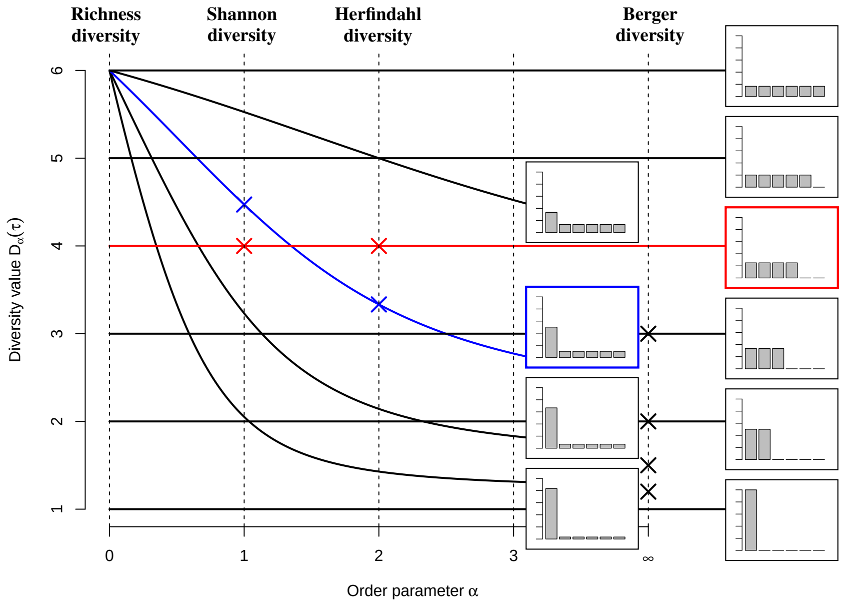


Figure 1: Values of different true diversities, depending on the order α, for different distributions.

2.4. *Relative true diversities*

As with the Rényi entropy, true diversities can be generalized to form a family of divergence measures. *Relative true diversities* generalize the family of true diversities by allowing to take any baseline other than the uniform distribution (that is, the distribution with maximal diversity). In different applications, it might be interesting to measure diversity with respect to another reference distribution. In Bayesian inference, for example, the divergence of the posterior, relative to the prior probability distribution is a measure of information gained. Relative true diversities generalize this notion using true diversity.

Note that this generalization is analogous to the well-known generalization from the family of *Rényi entropies* to the family of *Rényi divergences* [95, 96]. Among these generalizations, a well-known special case is the generalization from the *Shannon entropy* to the *Kullback-Leibler divergence* (also known as *relative entropy*) [97, 98].

Abusing notation, we also denote D_α the α-order relative true diversity between two distributions $p, q \in \Delta^{k-1}$, as described below.

Definition 3 (Relative true diversity) *The relative true diversity of order α is an application $D_\alpha : \Delta^* \times \Delta^* \to \mathbb{R}^+$, such that, given $p = (p_1, \ldots, p_k) \in \Delta^{k-1}$, $q = (q_1, \ldots, q_k) \in \Delta^{k-1}$, with $p_i = 0$ whenever $q_i = 0$, and $\alpha \in \mathbb{R}^+$. It is defined as*

$$D_\alpha(p \,\|\, q) = \left(\sum_{\substack{i=1 \\ q_i \neq 0}}^{k} p_i^\alpha q_i^{1-\alpha} \right)^{\frac{1}{\alpha-1}} \qquad \text{if } \alpha \neq 1,$$

As with true diversities, the extreme values are defined as the result of limit processes (cf. Theorems 4, 5, & 6 of [96]):

$$D_0(p \,\|\, q) := |\{i \in \{1, \ldots, k\} : p_i \neq 0 \text{ and } q_i \neq 0\}|,$$

$$D_1(p \,\|\, q) := \left(\prod_{\substack{i=1 \\ q_i \neq 0}}^{k} \left(\frac{p_i}{q_i}\right)^{p_i} \right)^{-1} \text{ with } {p_i}^{p_i} := 1 \text{ if } p_i = 0, \quad \text{and} \quad D_\infty(p \,\|\, q) := \left(\max_{\substack{i \le k \\ q_i \neq 0}} \frac{p_i}{q_i} \right)^{-1}.$$

This definition is analogous to that of true diversities with respect to the Rényi entropy: $D_\alpha(p \,\|\, q) = e^{H_\alpha(p \,\|\, q)}$. Thus, relative true diversities satisfy analogous properties. If $u = (1/k, \ldots, 1/k)$ is the uniform distribution, then, for $p \in \Delta^{k-1}$ we have that $D_\alpha(p \,\|\, u) = k/D_\alpha(p)$, and thus $D_\alpha(p \,\|\, u) \in [1, k]$ (1 when p is also uniform and k when $D_\alpha(p)$ is minimal, i.e., equal to 1). For a fixed k and a fixed $p \in \Delta^{k-1}$, relative true diversity is only minimal when distributions are equal. For all $p, q \in \Delta^{k-1}$

$$D_\alpha(p \,\|\, q) \geq D_\alpha(p \,\|\, p),$$

and its minimal value is $D_\alpha(p \,\|\, p) = 1$.

2.5. *Joint distributions, additivity, and Shannon entropy*

Other relevant properties of diversity measures are related to situations in which we have concurrent classifications. Following the notation from Section 2.1, let us consider a system in which items can be classified according to two criteria, giving rise to two relations: $\tau_1 \subseteq I \times T_1$ and $\tau_2 \subseteq I \times T_2$. For instance, books in a bookcase can be classified according to their genre (e.g. comics, novels) but also according to their author.

Let us define the *joint membership relation* $\tau_1{\times}\tau_2 \subseteq I \times (T_1 \times T_2)$ such that $(i, (t_1, t_2)) \in \tau_1{\times}\tau_2 \Leftrightarrow (i, t_1) \in \tau_1 \wedge (i, t_2) \in \tau_2$. Let us also define the *conditional membership relation* $(\tau_2 \mid t_1) \subseteq I \times T_2$ such that $(i, t_2) \in (\tau_2 \mid t_1) \Leftrightarrow (i, (t_1, t_2)) \in \tau_1{\times}\tau_2$.

As in Section 2.1, let us consider the following distributions: $p_{\tau_1}(t) = a_{\tau_1}(t)/|\tau_1|$ and $p_{\tau_2}(t) = a_{\tau_2}(t)/|\tau_2|$, resulting in $p_{\tau_1} \in \Delta^{|T_1|-1}$ and $p_{\tau_2} \in \Delta^{|T_2|-1}$. Similarly, we can define joint and conditional distributions. We define the *joint distribution* over T_1 and T_2 as

$$p_{\tau_1\times\tau_2}(t) = \frac{a_{\tau_1\times\tau_2}(t)}{|\tau_1 \times \tau_2|}, \quad \text{with } p_{\tau_1\times\tau_2} \in \Delta^{(|T_1|-1)(|T_2|-1)},$$

and the *conditional distribution* over T_2 given $t_1 \in T_1$ as

$$p_{(\tau_2 \mid t_1)}(t) = \frac{a_{(\tau_2 \mid t_1)}(t)}{|(\tau_2 \mid t_1)|}, \quad \text{for } t_1 \in T_1, \text{ with } p_{(\tau_2 \mid t_1)} \in \Delta^{|T_2|-1}.$$

The first of two additivity principles to be considered in this article is the *weak additivity principle*.

Definition 4 (Weak additivity) *A diversity measure D is said to satisfy the principle of weak additivity if and only if, for all τ_1 and τ_2 such that $p_{\tau_1\times\tau_2}(t_1, t_2) = p_{\tau_1}(t_1)p_{\tau_2}(t_2)$, we have $D(p_{\tau_1\times\tau_2}) = D(p_{\tau_1})D(p_{\tau_2})$.*

In other words, if two classifications are independent, then the diversity of the joint classification is equal to the product of the diversities of each separate one.

Theorem 6 (True diversities satisfy the principle of weak additivity [81]) *True diversities D_α satisfy the principle of weak additivity.*

Note that Theorem 6 is equivalent to the expression of the joint Rényi entropy for independent variables.

A stronger property, called *strong additivity principle*, and not restricted to independence between τ_1 and τ_2, is verified for the particular case of the 1-order true diversity, or Shannon diversity.

Definition 5 (Strong additivity) *A diversity measure D is said to satisfy the strong additivity principle if and only if, for all τ_1 and τ_2, we have $D(p_{\tau_1\times\tau_2}) = D(p_{\tau_1})D(p_{\tau_2 \mid \tau_1})$ where $D(p_{\tau_2 \mid \tau_1}) = \prod_{t_1 \in T_1} D(p_{\tau_2 \mid t_1})^{p_{\tau_1}(t_1)}$.*

In other words, the diversity of the joint classification is equal to the diversity of the first classification times the diversity of the second classification conditioned by the knowledge of the first one. Note that *conditional diversity* is the weighted geometric mean of the diversities of conditional distributions.

Theorem 7 (1-order true diversity satisfies the strong additivity principle [81]) *1-order true diversity D_1 satisfies the principle of strong additivity.*

The principle of strong additivity is analogous to the well-known *chain rule* between *conditional entropy* and *joint entropy* in information theory (cf. Section 2.5 [98]): $H(X, Y) = H(X) + H(Y|X)$ for random variables X and Y.

3. Random Walks in Heterogeneous Information Networks

In the previous section, we presented a broad definition of diversity, which we then narrowed to a particular family of measures that share relevant properties captured by axioms. The functions of the theory determined by these axioms resulted in the true diversities, which were shown to be connected to many of the diversity measures used in different domains of research.

When considering complex systems and network-structured data, the concept of diversity can designate different observables of a system. This is related not only to the way in which diversity is measured from a distribution, but also to the way different distributions can be computed from the system. In this article, we develop a single framework for both operations, effectively covering and summarizing the measurement of diversity in networks in several domains. In order to do so, we develop in this section a formalism for the treatment of networks that are relevant for fields interested in diversity.

Developed within graph theory, heterogeneous information networks [24, 25] (equivalent to directed graphs with colored vertices and edges) have been recently used to provide ontologies to represent complex unstructured data in a wide gamut of applications (knowledge graphs are prominent examples their flexibility [99, 100, 101]). For this work, we will consider an extended model, using multigraphs (graphs for which multiple edges might exist between any given couple of vertices), for the development of a framework for measuring diversity in networks. As we shall see in more details in Section 5, many situations encountered in practice can be represented using heterogeneous information networks. For example, when modeling the consumption of news on a website, the situation may be represented as users selecting articles, and these articles having specific categories (business, culture, sports, etc.). This translates to a heterogeneous information network with three vertex types (users, articles, categories) and two edge types (users select articles, articles belong to categories). Section 5 provides further, more detailed examples.

3.1. Preliminary notations

In this article we consider a multigraph composed of a set of nodes $\boldsymbol{V}$, linked by a set of directed edges $\boldsymbol{E}$. We propose the following system of capitalization and typefaces to referentiate different objects:

- *vertices* and *edges* are designated by lowercase letters, v and e;
- a set of *types of vertex* is designated by $\mathcal{A}$;
- a set of *types of edge* is designated by $\mathcal{R}$;
- types in $\mathcal{A}$ are notated with uppercase letters A, types in $\mathcal{R}$ are denoted with uppercase letters R;
- *vertex sets* and *edge sets* by uppercase letters V and E;
- sets of *vertex types* and *edge types* labels by calligraphic letters $\mathcal{V}$ and $\mathcal{E}$;
- random variables with support on sets of vertices by the capital letter X.

3.2. Heterogeneous information networks

In contrast to traditional formalizations of heterogeneous information networks [24, 25], we propose the use of multigraphs as a basis, for generality. A *multigraph* G is a couple $(\boldsymbol{V}, \boldsymbol{E})$ where $\boldsymbol{V} = \{v_1, \ldots, v_n\}$ is a set of vertices and $\boldsymbol{E} = \{e_1, \ldots, e_m\}$ is a set of directed edges that is a multisubset of $\boldsymbol{V} \times \boldsymbol{V}$. Given an edge $e \in \boldsymbol{E}$, we denote $v_{\text{src}}(e)$ its source vertex and $v_{\text{dst}}(e)$ its destination vertex such that $(v_{\text{src}}(e), v_{\text{dst}}(e)) \in \boldsymbol{V} \times \boldsymbol{V}$.

We also denote $\epsilon : \boldsymbol{V} \times \boldsymbol{V} \to \mathbb{N}$ the multiplicity function of the edges, that is, the function counting the number of edges in $\boldsymbol{E}$ that link any two vertices: $\epsilon(v_1, v_2) = |\{e \in \boldsymbol{E} : v_{\text{src}}(e) = v_1 \wedge v_{\text{dst}}(e) = v_2\}|$. We also define:

- $\epsilon(v_1, -) \coloneqq \sum_{v_2 \in \boldsymbol{V}} \epsilon(v_1, v_2)$ the *out-degree* of vertex v_1;
- $\epsilon(-, v_2) \coloneqq \sum_{v_1 \in \boldsymbol{V}} \epsilon(v_1, v_2)$ the *in-degree* of vertex v_2;
- $\epsilon(-, -) \coloneqq \sum_{(v_1, v_2) \in \boldsymbol{V} \times \boldsymbol{V}} \epsilon(v_1, v_2)$ the *total number* of edges.

We now define heterogeneous information networks using multigraphs. Being more general, this allows for the representation of a wider gamut of situations. Classical heterogeneous information networks can be easily accounted for by constraining the multiplicity of edges.

Definition 6 (Heterogeneous information network) *A* heterogeneous information network $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \mathcal{A}, \mathcal{R}, \varphi, \psi)$ *is a multigraph* $(\boldsymbol{V}, \boldsymbol{E})$*, with a vertex labeling function* $\varphi : \boldsymbol{V} \to \mathcal{A}$ *and an edge labeling function* $\psi : \boldsymbol{E} \to \mathcal{R}$*, such that edges having the same type in* $\mathcal{R}$ *have their source vertices mapped to the same type in* $\mathcal{A}$ *and their destination vertices mapped to the same type in* $\mathcal{A}$*:*

$$\forall\, e, e' \in \boldsymbol{E} \ \ (\ \ \psi(e) = \psi(e') \quad \Rightarrow \quad (\ \varphi(v_{src}(e)) = \varphi(v_{src}(e')) \quad \wedge \quad \varphi(v_{dst}(e)) = \varphi(v_{dst}(e'))\ )\ \ ).$$

Labeling functions φ and ψ, mapping vertices to vertex types and edges to edge types, induce a partition in the set of vertices and a partition in the set of edges. If $\mathcal{A} = \{A_1, \ldots, A_N\}$ and $\mathcal{R} = \{R_1, \ldots, R_M\}$, φ and ψ induce partitions $\mathcal{V} = \{V_1, \ldots, V_N\}$ on $\boldsymbol{V}$ and $\mathcal{E} = \{E_1, \ldots, E_M\}$ on $\boldsymbol{E}$. These partitions are such that $\forall v \in \boldsymbol{V}$ $(\varphi(v) = A_i \Leftrightarrow v \in V_i)$ and $\forall e \in \boldsymbol{V}$, $(\varphi(e) = R_j \Leftrightarrow e \in E_j)$. Thus, abusing notation, we make indistinct use of types in $\mathcal{A}$ and sets in $\mathcal{V}$, and of types in $\mathcal{R}$ and sets in $\mathcal{E}$ when this is not ambiguous.

Given an edge type $E \in \mathcal{E}$, we denote $V_{\text{src}}(E) \in \mathcal{V}$ its source-vertex type and $V_{\text{dst}}(E) \in \mathcal{V}$ its destination-vertex type. We also denote $\epsilon_E : V_{\text{src}}(E) \times V_{\text{dst}}(E) \to \mathbb{N}$ the specialization of ϵ on E, that is, the function counting the number of edges in E going from a given vertex in $V_{\text{src}}(E)$ to a given vertex in $V_{\text{dst}}(E)$:

$$\epsilon_E(v_1, v_2) = |\{e \in E : v_{\text{src}}(e) = v_1 \wedge v_{\text{dst}}(e) = v_2\}|.$$

As before, we also define:

- $\epsilon_E(v_1, -) \coloneqq \sum_{v_2 \in V_{\text{dst}}(E)} \epsilon_E(v_1, v_2)$ is the *out-degree* of v_1 among edges in E;
- $\epsilon_E(-, v_2) \coloneqq \sum_{v_1 \in V_{\text{src}}(E)} \epsilon_E(v_1, v_2)$ is the *in-degree* of v_2 among edges in E;
- $\epsilon_E(-, -) \coloneqq \sum_{(v_1, v_2) \in V_{\text{src}}(E) \times V_{\text{dst}}(E)} \epsilon_E(v_1, v_2)$ is the *number* of edges in E.

Following the example of existing definitions for heterogeneous information networks [24, 25, 102], we define the network schema. Consistency in the direction of edges belonging to the same edge type allows for its definition as a proper directed graph. Figure 2 illustrates a heterogeneous information network and its network schema.

Definition 7 (Network schema) *The network schema of a heterogeneous information network* $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \mathcal{A}, \mathcal{R}, \varphi, \psi)$ *is the directed graph* $\mathcal{S} = (\mathcal{E}, \mathcal{V})$ *that has vertex types* $\mathcal{V}$ *for vertices and edge types* $\mathcal{E}$ *for edges.*

Let us now define the probability of transitioning between vertices randomly following the available directed edges from an edge type.

Definition 8 (Probability of transitioning between vertices in an edge type) *Given an edge type $E \in \mathcal{E}$, assuming that each vertex in $V_{src}(E)$ is connected to at least one vertex in $V_{dst}(E)$, i.e. $\forall v_1 \in V_{src}(E)$ $(\epsilon_E(v_1, -) > 0)$, we denote by $p_E : V_{src}(E) \times V_{dst}(E) \to [0, 1]$ the* transition probability *of the random walk following edges in E, for all $(v_1, v_2) \in V_{src}(E) \times V_{dst}(E)$, as*

$$p_E(v_2 \mid v_1) := \frac{\epsilon_E(v_1, v_2)}{\epsilon_E(v_1, -)}.$$

Definition 9 (Random transition between vertices in an edge type) *For an edge type $E \in \mathcal{E}$ going from vertex type $V_{src}(E)$ to vertex type $V_{dst}(E)$ in $\mathcal{V}$, we denote the transition from a random vertex $X_{src} \in V_{src}(E)$ to a random vertex $X_{dst} \in V_{dst}(E)$, following probability distribution p_E, as $X_{src} \xrightarrow{E} X_{dst}$.*

As a consequence of Definition 8, $\forall v_1 \in V_{\mathrm{src}}(E),\ p_E(\cdot \mid v_1) : V_{\mathrm{dst}}(E) \to \mathbb{R}^+$ is a probability distribution on $V_{\mathrm{dst}}(E)$. For all $v_2 \in V_{\mathrm{dst}}(E)$ we have that $p_E(v_2 \mid v_1) \in [0, 1]$ and $\sum_{v_2 \in V_{\mathrm{dst}}(E)} p_E(v_2 \mid v_1) = 1$.

In the case where vertex $v_1 \in V_{\mathrm{src}}(E)$ is connected to no vertex in $V_{\mathrm{dst}}(E)$ (i.e. when $\epsilon_E(v_1, -) = 0$), $p_E(v_2 \mid v_1)$ cannot be defined as above. This situation can be remedied by adding a sink vertex to each vertex type. For every $E \in \mathcal{E}$, an edge e^s_E is added such that $v_{\mathrm{src}}(e^s_E)$ is the sink vertex in $V_{\mathrm{src}}(E)$, and such that $v_{\mathrm{dst}}(e^s_E)$ is the sink vertex in $V_{\mathrm{dst}}(E)$. Then, vertices in $V_{\mathrm{src}}(E)$ connected to no vertex in $V_{\mathrm{dst}}(E)$ can be connected to the sink vertex. In the rest of this article, we will assume that this procedure has been applied if needed and that for every $E \in \mathcal{E}$, there are no vertices in $V_{\mathrm{src}}(E)$ that are not connected to at least one vertex in $V_{\mathrm{dst}}(E)$.

3.3. Meta paths and constrained random walks

Random walks in heterogeneous information networks can be constrained [34, 103] to follow a specific sequence of edge types, called *meta path* [102, 104]. This allows for the computation of the probability distribution of the ending vertex of a random walker constrained to a specific meta path. The variety and combinatorics of meta paths is what will give birth, in next section, to the network diversity measures that this article proposes.

For the definition of meta paths, we will first consider sequences on the set edge types. A sequence of length k for $M \geq 1$ is a k-tuple $r = (r_1, \ldots, r_k)$ such that for all $i \in \{1, \ldots, k\}$ we have that $r_i \in \{1, \ldots, M\}$.

Definition 10 (Meta path) *Given a heterogeneous information network $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \mathcal{A}, \mathcal{R}, \varphi, \psi)$ and a sequence r of length $k \in \mathbb{N}$ for $M = |\mathcal{R}|$, a* meta path *of length k is the k-tuple $\Pi = (E_{r_1}, \ldots, E_{r_k}) \in \mathcal{E}^k$ of k edge types (with possible repetitions) such that the source vertex type of an edge type is the destination vertex type of the previous one in the k-tuple Π: i.e. $\forall 1 \leq i \leq k,\ V_{src}(E_{r_i}) = V_{dst}(E_{r_{i-1}})$.*

We denote by $V_{\mathrm{src}}(\Pi) = V_{\mathrm{src}}(E_{r_1})$ the source vertex type of path type Π, and by $V_{\mathrm{dst}}(\Pi) = V_{\mathrm{dst}}(E_{r_k})$ its destination vertex type. Figure 2 provides an illustration of a heterogeneous information network and a meta path on its network schema.

Using the notion of meta path, we define a random walk restricted to it.

Definition 11 (Random walk constrained to a meta path) *Given a meta path $\Pi = (E_{r_1}, \ldots, E_{r_k})$ of length k and a random variable $X_0 \in V_{src}(\Pi)$ representing the starting position of a random walk in vertex type $V_{src}(\Pi)$, the* associated random walk restricted to Π *is a sequence of $k + 1$ random variables $(X_0, X_1, \ldots, X_k)$ resulting from the sequential random transition between vertices in the edge types (cf. Definition 8) of Π:*

$$X_0 \xrightarrow{E_{r_1}} X_1 \xrightarrow{E_{r_2}} X_2 \xrightarrow{E_{r_3}} \cdots \xrightarrow{E_{r_k}} X_k,$$

where, for all i, $X_i \in V_{dst}(E_{r_i})$.

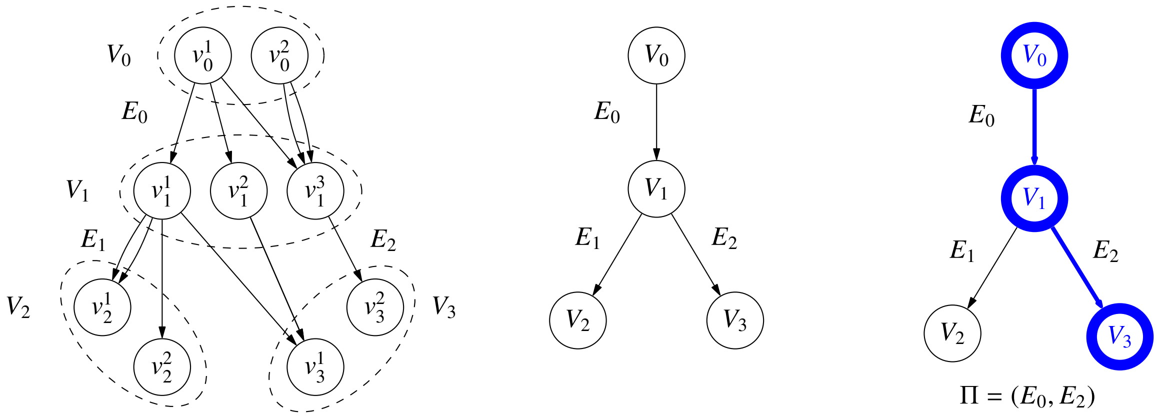


Figure 2: A heterogeneous information network (left), its network schema (center), and a meta path Π on the network schema (right).

This is known as path-constrained random walk in the information retrieval community [34, 103].

It follows from Definitions 8 and 11 that a random walk restricted to a meta path Π of length k is a Markov chain with transition probabilities defined as

$$\Pr(X_i = v_i \mid X_{i-1} = v_{i-1}) = P_{E_{r_i}}(v_i \mid v_{i-1}),$$

for $v_{i-1} \in V_{\text{src}}(E_{r_i})$ and $v_i \in V_{\text{dst}}(E_{r_i})$.

For the next two definitions we consider a meta path $\Pi = (E_{r_1}, \ldots, E_{r_k})$ of length k and its associated random walk restricted to Π, i.e., the sequence $(X_0, X_1, \ldots, X_k)$ of random variables. The probability distribution for the ending vertex of the random walk in $V_{\text{dst}}(\Pi)$ plays an central role in the network diversity measures that will be proposed in the next section. Let us define the conditional and the unconditional probability distributions.

Definition 12 (Conditional probability distribution for random walks) *The conditional probability distribution of $X_k \in V_{dst}(\Pi)$, the destination vertex of the random walk constrained to Π, given that it started in $v_0 \in V_{src}(\Pi)$ (i.e. $X_0 = v_0$) is denoted by $p_\Pi(v_k \mid v_0)$ for $v_k \in V_{dst}(\Pi)$, and can be computed recursively as:*

$$\begin{aligned} p_\Pi(v_k \mid v_0) &= \Pr(X_k = v_k \mid X_0 = v_0) \\ &= \sum_{v_0 \in V_{src}(E_{r_1})} p_{(E_{r_2},\ldots,E_{r_k})}(v_k \mid v_1)\ p_{E_{r_1}}(v_1|v_0). \end{aligned}$$

We will also designate by $p_{\Pi|v_0}(v_k)$ the distribution $p_\Pi(v_k|v_0)$ over the vertices of $V_{dst}(\Pi)$.

Using the conditional probability distribution, its unconditional version can be computed.

Definition 13 (Unconditional probability distribution for random walks) *The unconditional probability distribution of $X_k \in V_{dst}(\Pi)$, the destination vertex of the random walk restrained to Π, is denoted by $p_\Pi(v_k)$ for $v_k \in V_{dst}(\Pi)$, and can be computed applying the law of total probability to the conditional distribution $p_{\Pi \mid v_0}$ as:*

$$\begin{aligned} p_\Pi(v_k) &= \Pr(X_k = v_k) \\ &= \sum_{v_0 \in V_{src}(\Pi)} p_{\Pi \mid v_0}(v_k)\ \Pr(X_0 = v_0). \end{aligned}$$

In Definition 13, the dependence of p_Π on $\Pr(X_0 = v_0)$ (the probability distribution for the starting vertex) is seen explicitly.

We now consider the edges resulting from the projection of all edge types in a meta path Π. This operation, related to the counting of paths in meta paths, is used in the literature in related measures. The counting of the number of paths between vertices in a heterogeneous information networks is used, for example, in the construction of similarity metrics for vertex search [33] and in recommender systems [27].

Definition 14 (Projection of a meta path) *Given a meta path $\Pi = (E_{r_1}, \ldots, E_{r_k})$, we denote by E_Π the set of edges going from vertices in $V_{src}(\Pi)$ to vertices in $V_{dst}(\Pi)$, and resulting from the projection of all paths in the meta path Π. We denote $\epsilon_\Pi(v_0, v_k)$ the number of paths starting at $v_0 \in V_{src}(\Pi)$ and ending at $v_k \in V_{dst}(\Pi)$, that are part of meta path Π. It is recursively computed as*

$$\epsilon_{E_\Pi}(v_0, v_k) \quad = \sum_{v_1 \in V_{dst}(E_{r_1})} \epsilon_{E_{r_1}}(v_0, v_1)\ \epsilon_{(E_{r_1},\ldots,E_{r_k})}(v_2, v_k),$$

with $\epsilon_{(E_{r_k}, E_{r_k})} = \epsilon_{E_{r_k}}$.

The projection is such that there is an edge in E_Π for each path in Π. This allows for the definition of a –one step– random walk from $V_{\text{src}}(\Pi)$ to $V_{\text{dst}}(\Pi)$. Its probability distribution is denoted p_{E_Π}, and computed following Definition 8. If random walk $X_0 \xrightarrow{E_{r_1}} X_1 \xrightarrow{E_{r_2}} \cdots \xrightarrow{E_{r_k}} X_k$ involves choosing a random edge at each step at each vertex type $V_{\text{src}}(E_{r_i})$, the random walk $X_0 \xrightarrow{E_\Pi} X_k$ involves choosing randomly one path among all those possible in Π.

4. Network Diversity Measures

In the previous section we established a formal framework for heterogeneous information networks within which we defined meta paths and random walks constrained to them. This allowed us to consider different probability distributions related to these random walks. In this section, we apply true diversity measures to these distributions, completing the framework for the measurement of diversity in heterogeneous information networks.

Depending on the chosen meta path, one can compute several diversities in a network. These diversities will correspond to different concepts related to the structure of the vertices and edges in the meta path: *individual*, *collective*, *relative*, *projected*, and *backward* diversity (to be defined in this section). These concepts will find in turn different semantical contents depending on what is being modeled by the heterogeneous information network. The way in which diversities associated with meta paths can correspond to different concepts will be made clear in this section, and illustrated through different applications in the next section.

All the definitions and results in this section refer to an heterogeneous information network $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E}, \mathcal{A}, \mathcal{R}, \varphi, \psi)$, and a meta path $\Pi = (E_{r_1}, \ldots, E_{r_k})$ of length k going from vertex type $V_{\text{src}}(\Pi)$ to vertex type $V_{\text{dst}}(\Pi)$. In the scope of this section, let us define $V_{\text{start}} = V_{\text{src}}(\Pi)$ and $V_{\text{end}} = V_{\text{dst}}(\Pi)$ for ease of notation. For diversities defined in this section, we will talk about the diversity of a given vertex type with respect to another one on a heterogeneous information network, and along a given meta path. In other words, given $\mathcal{G}$, in this section we define the network diversities for a given Π, that are of the form: *V_{end} diversity of V_{start} along* Π.

4.1. Collective and individual diversity

The collective V_{end} diversity of vertices V_{start} along meta path Π is the diversity of the probability distribution on the vertices of V_{end} resulting from a random walk, starting at a random vertex on V_{start}, and restricted to meta path Π. Using the previous definitions, we formally define this quantity.

Definition 15 (Collective diversity) *Given the probability distribution* $\Pr(X_0)$ *of starting at a random starting vertex $X_0 \in V_{start}$, we define the* collective V_{end} diversity of V_{start} along Π *as the true diversity of the probability distribution of the ending vertex of the constrained random walk. We denote it as $D_\alpha\left(X_0 \xrightarrow{\Pi} X_k\right)$ and compute it as*

$$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k\right) \quad = \quad D_\alpha(p_\Pi).$$

Note that this measure depends on the starting probability distribution $\Pr(X_0 = v_0)$ and on transition probabilities $p_{E_{r_i}}(v_i \mid v_{i-1})$ for each $E_{r_i} \in \Pi$. Figure 3 provides an example of the measurement of the collective diversity for a simple heterogeneous information network containing 5 vertices (represented as circles) and 6 edges (represented as arrows between circles), and using two different starting probabilities distributions $\Pr(X_0)$. In Figure 3, vertices are organized into two vertex types $V_0 = \{v_0^1, v_0^2\}$ and $V_1 = \{v_1^1, v_1^2, v_1^3\}$ (represented as two horizontal layers) and edges are

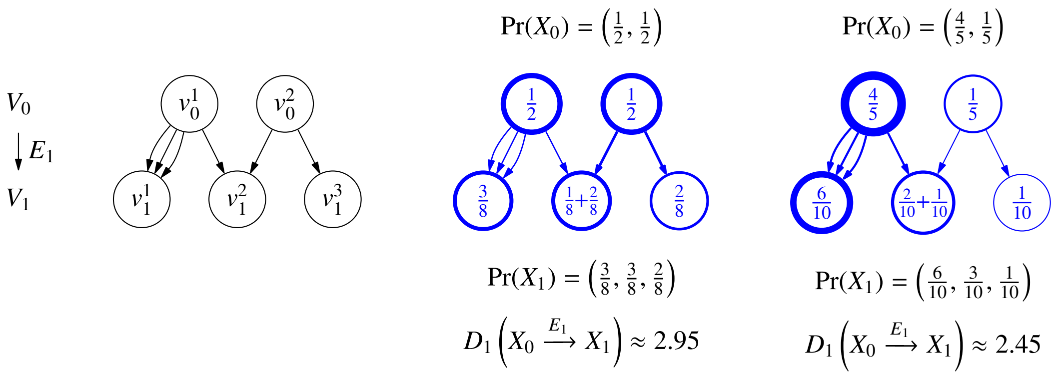


Figure 3: Computation of the collective V_1 diversity of V_0 along a simple meta path made of only one edge type. Diversity along a path type depends on the starting probability distribution $\Pr(X_0)$ and on the transition probabilities.

organized into a unique edge type E_1, going from V_0 to V_1. Two examples of measurements are illustrated in blue for two different starting distributions (numbers within circles give the probabilities for the starting position of random walk during the different steps of the walk).

The choice of $X_0 \sim$Uniform(V_{start}) has a central role in applications, because by giving each node on V_{start} equal chance of being the start of the random walk, the resulting collective diversity will be that of the collective –equal– contribution of nodes in V_{start}. Similarly, considering a subset $V'_{\text{start}} \subset V_{\text{start}}$ and choosing $X_0 \sim$Uniform(V'_{start}) allows for the collective diversity of the group of nodes V'_{start}.

It will be argued that the *conditioned* probabilities of random walks along a path Π are also of interest, as they convey information about the structure of the network reachable from some vertices in V_{start}. In particular, given a starting vertex $v_0 \in V_{\text{start}}$, we define the *individual V_{end} diversity of v_0 along* Π as the true diversity of the probability distribution of $X_k \in V_{\text{end}}$ for the end of the constrained random walk, knowing that it started at a given $v_0 \in V_{\text{start}}$ (i.e, $X_0 = v_0$). Figure 4 illustrates the measurement of the individual diversity for two different vertices in V_{start} for a simple heterogeneous information network.

Definition 16 (Individual diversity) *Given a starting vertex $v_0 \in V_{start}$, we define the* individual V_{end} diversity of v_0 along Π *as the true diversity of the probability distribution of the ending vertex of the constrained random walk. We denote it as $D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0 = v_0\right)$ and compute it as*

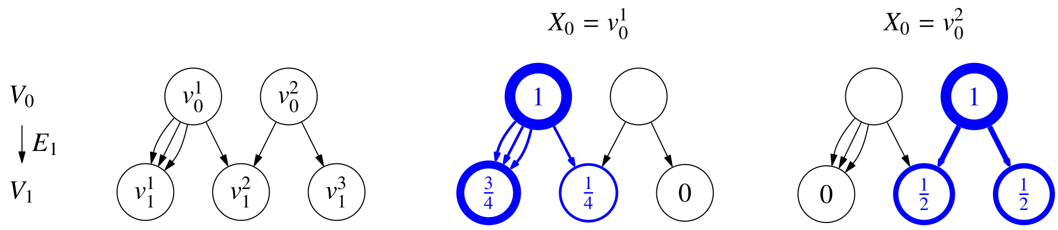
$$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0 = v_0\right) \quad = \quad D_\alpha(p_{\Pi|v_0}).$$

An aggregation of individual diversities can be computed to represent the diversities of all (or many) of the vertices in starting vertex type V_{start}. Following the definition of conditional entropy in information theory (cf. Section 2.2 in [98]), we define the *mean V_{end} individual diversity of V_{start} along* Π as the weighted geometric mean of individual diversities.

Definition 17 (Mean individual diversity) *Given a starting vertex $v_0 \in V_{start}$, we define the* mean individual V_{end} diversity of v_0 along Π *as the weighted geometric mean of the individual diversities. We denote it by $D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0\right)$ and compute it as*

$$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0\right) \quad = \quad \prod_{v_0 \in V_{start}} D_\alpha(p_{\Pi|v_0})^{\Pr(X_0=v_0)}$$

Note that this mean is weighted by –and so depend on– the starting probability distribution $\Pr(X_0)$ over V_0, and that it is minimal (i.e., equal to 1) when each individual diversity is minimal. The mean individual diversity is a weighted geometric mean in the general case (i.e., for any distribution for X_0), and a geometric mean when any vertex in V_{start} has the same probability of being the starting point of the random walk (i.e., when $X_0 \sim \text{Uniform}(V_{\text{start}})$). As



$$\Pr(X_1 \mid X_0 = v_0^1) = \left(\tfrac{3}{4}, \tfrac{1}{4}, 0\right) \qquad \Pr(X_1 \mid X_0 = v_0^2) = \left(0, \tfrac{1}{2}, \tfrac{1}{2}\right)$$

$$D_1\left(X_0 \xrightarrow{E_1} X_1 \mid X_0 = v_0^1\right) \approx 1.75 \quad D_1\left(X_0 \xrightarrow{E_1} X_1 \mid X_0 = v_0^2\right) = 2$$

$$\Pr(X_0) = \left(\tfrac{1}{2}, \tfrac{1}{2}\right) \quad \Rightarrow \quad D_1\left(X_0 \xrightarrow{E_1} X_1 \mid X_0\right) \approx (1.75)^{\frac{1}{2}}\ (2)^{\frac{1}{2}} \approx 1.87$$

$$\Pr(X_0) = \left(\tfrac{4}{5}, \tfrac{1}{5}\right) \quad \Rightarrow \quad D_1\left(X_0 \xrightarrow{E_1} X_1 \mid X_0\right) \approx (1.75)^{\frac{4}{5}}\ (2)^{\frac{1}{5}} \approx 1.80$$

Figure 4: Examples of a heterogeneous information network (top left), the individual diversities of two vertices (top center and top right), and the mean individual diversities for two different starting probability distributions (bottom).

with the collective diversity, a uniform mean individual diversity of groups of vertices $V'_{\text{start}} \subset V_{\text{start}}$ can be considered choosing $V'_{\text{start}} \sim \text{Uniform}(V'_{\text{start}})$. Figure 4 illustrates the computation of individual and mean individual diversities in a simple heterogeneous information network.

Individual and collective diversities are two complementary measures describing different properties of the system, as illustrated in Figure 5. It is possible for a system to have a low mean individual diversity while having a high collective diversity (top-right in Figure 5), or a high mean individual diversity while having a low collective diversity (bottom-left in Figure 5).

4.2. Backward diversity

Backward diversity is related to random walks following directions and edges opposite to those of a given meta path. In order to treat them formally, we first present the following definitions.

Definition 18 (Transpose edge type) *Let $E \in \mathcal{E}$ be an edge type. We denote by $E^\intercal$ the set of edges resulting from inverting those of E:*

$$E^\intercal = \{(v_{dst}(e), v_{src}(e)) \in \boldsymbol{V} \times \boldsymbol{V} : e \in E\}\,.$$

Definition 19 (Transpose meta path) *For a meta path $\Pi = (E_{r_1}, \ldots, E_{r_k})$ we define its transpose meta path $\Pi^\intercal$ as*

$$\Pi^\intercal = (E_{r_k}^\intercal, \ldots, E_{r_1}^\intercal).$$

Using random walks along a meta path Π, we can also compute a distribution on V_{start} as the probability of the starting point of the constrained random walk at a given vertex $X_0 \in V_{\text{start}}$ when the ending vertex $v_k \in V_{\text{end}}$ is known. The true diversity of this distribution, called the *backward V_{start} diversity of $v_k \in V_{end}$ along* Π, provides a value for the diversity of starting points that can reach v_k following Π.

Definition 20 (Backward diversity) *Given an ending vertex $v_k \in V_{end}$, we define the* individual V_{end} backward diversity of V_{start} along Π *as the true diversity of the distribution of the starting vertex $X_0 \in V_{start}$. We denote it by $D_\alpha\left(X_0 \mid X_0 \xrightarrow{\Pi} X_k = v_k\right)$ and compute it as*

$$D_\alpha\left(X_0 \mid X_0 \xrightarrow{\Pi} X_k = v_k\right) \;=\; D_\alpha\left(p_{\Pi^\intercal|v_k}\right).$$

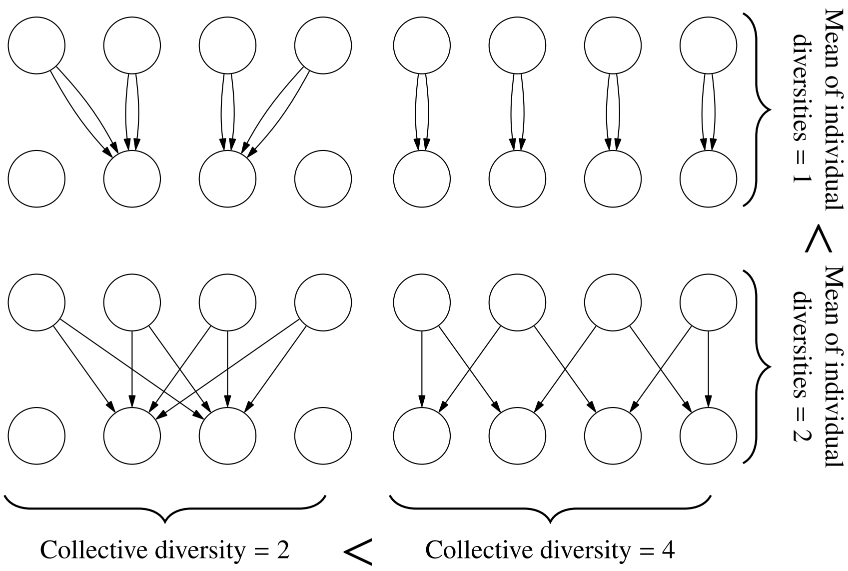


Figure 5: Different heterogeneous information network with two vertex types, illustrating different relative ordering for collective and mean individual diversity.

We denote by $D_\alpha\left(X_0 \mid X_0 \xrightarrow{\Pi} X_k\right)$ *the mean backward diversity and compute it as*

$$D_\alpha\left(X_0 \mid X_0 \xrightarrow{\Pi} X_k\right) = \prod_{v_k \in V_{end}} D_\alpha\left(p_{\Pi^\intercal|v_k}\right)^{\Pr(X_k=v_k)}.$$

Collective and individual diversities are not always independent for all true diversities, as evidenced by the following result involving backward diversities, for which we need first to be able to consider parts of meta paths.

Definition 21 (Parts of meta paths) *Given a meta path* $\Pi = (E_{r_1}, \ldots, E_{r_k})$ *of length k we denote by* $\Pi_{(i,j)}$*, for* $1 \leq i \leq j \leq k$*, its restriction*

$$\Pi_{(i,j)} = \left(E_{r_i}, E_{r_{i+1}}, \ldots E_{r_{j-1}}, E_{r_j}\right).$$

Theorem 8 (Bound for collective Shannon diversity) *The following inequality holds for the Shannon, or 1-order true diversity,*

$$D_1(X_0 \xrightarrow{\Pi} X_k) \quad \leq \quad D_1(X_0 \xrightarrow{\Pi_{(1,i)}} X_i)\, D_1(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k \mid X_0 \xrightarrow{\Pi_{(1,i)}} X_i),$$

with equality if and only if $D_1(X_0 \xrightarrow{\Pi_{(1,i)}} X_i \mid X_i \xrightarrow{\Pi_{(i+1,k)}} X_k) = 1$.

In other words, the collective diversity along a meta path is bounded by two factors: (1) the collective diversity at any step of the meta path, multiplied by (2) the mean individual diversity along the remaining part of the meta path. This bound is achieved if and only if all individual diversities along a meta path are minimal.

Before proving Theorem 8, let us first prove the following result relating collective and individual 1-order true diversity for a single edge type.

Lemma 1 (Relation between collective, backward, and mean individual diversities) *Let us consider an edge type* E*, with* $V_{src}(E) = V_0$ *and* $V_{dst}(E) = V_1$*, and the constrained random walk* $X_0 \xrightarrow{E} X_1$*, with* $X_0 \in V_0$ *and* $X_1 \in V_1$*. The following identity relation between collective, backward, and mean individual 1-order true diversities holds:*

$$\underbrace{D_1\left(X_0 \xrightarrow{E} X_1\right)}_{collective} \underbrace{D_1\left(X_0 \mid X_0 \xrightarrow{E} X_1\right)}_{mean\ backward} = D_1\left(\Pr(X_0)\right) \underbrace{D_1\left(X_0 \xrightarrow{E} X_1 \mid X_0\right)}_{mean\ individual},$$

where $D_1(\Pr(X_0))$ is the 1-order true diversity of the distribution for the starting vertex of the random walk.

Proof. Let us consider the 1-order true diversity of the joint probability $\Pr(X_0, X_1)$ of the starting vertex on V_0 and the ending vertex on V_1. Despite X_0 and X_1 being dependent, by the principle of strong additivity of the 1-order true diversity (cf. Theorem 7), we have

$$\begin{aligned} D_1(\Pr(X_0, X_1)) &\stackrel{\text{Thm. 7}}{=} D_1(\Pr(X_0)) \prod_{v_0 \in V_0} D_1(\Pr(X_1 \mid X_0 = v_0))^{\Pr(X_0 = v_0)} \\ &\stackrel{\text{Def. 17}}{=} D_1(\Pr(X_0))\, D_1\left(X_0 \xrightarrow{E} X_1 \mid X_0\right). \end{aligned}$$

Also by the principle of strong additivity of the 1-order true diversity we have

$$\begin{aligned} D_1(\Pr(X_0, X_1)) &\stackrel{\text{Thm. 7}}{=} D_1(\Pr(X_1)) \prod_{v_1 \in V_1} D_1(\Pr(X_0 \mid X_1 = v_1))^{\Pr(X_1 = v_1)} \\ &\stackrel{\text{Def. 20}}{=} D_1\left(X_0 \xrightarrow{E} X_1\right) D_1\left(X_0 \mid X_0 \xrightarrow{E} X_1\right). \quad \blacksquare \end{aligned}$$

Since true diversities are greater or equal to 1 (cf. Theorem 3), it is clear that

$$D_1\left(X_0 \xrightarrow{E} X_1\right) \leq D_1(\Pr(X_0))\, D_1\left(X_0 \xrightarrow{E} X_1 \mid X_0\right),$$

with equality when the mean backward diversity is minimal, $D_1\left(X_0 \mid X_0 \xrightarrow{E} X_1\right) = 1$. This can only happen when each ending vertex in V_1 is reachable from only one starting vertex in V_0.

Using the same procedure as in Lemma 1, we can now prove Theorem 8.

Proof of Theorem 8. Given a meta path $\Pi = (E_{r_1}, \ldots, E_{r_k})$ and a constrained random walk $X_0 \xrightarrow{\Pi} X_k$ along it, let us split it in two parts, dividing our walk in two parts:

$$\Pi_{(1,i)} = (E_{r_1}, \ldots, E_{r_i}), \quad \text{for random walk } X_0 \xrightarrow{\Pi_{(1,i)}} X_i, \quad \text{and}$$

$$\Pi_{(i+1,k)} = (E_{r_{i+1}}, \ldots, E_{r_k}), \quad \text{for random walk } X_i \xrightarrow{\Pi_{(i+1,k)}} X_k.$$

Following the same argument from the proof of Lemma 1, we compute the 1-order diversity of the probability $\Pr(X_i, X_k)$, using the strong additivity principle to obtain two different expressions.

The first application of the strong additivity yields

$$D_1(\Pr(X_i, X_k)) = D_1\left(X_0 \xrightarrow{\Pi_{(1,i)}} X_i\right) D_1\left(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k \mid X_0 \xrightarrow{\Pi_{(1,i)}} X_i\right),$$

where $D_1\left(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k \mid X_0 \xrightarrow{\Pi_{(1,i)}} X_i\right)$ is the mean individual diversity along meta path $\Pi_{(i+1,k)}$ using the probabilities resulting from random walk $X_0 \xrightarrow{\Pi_{(1,i)}} X_k$ for the weighted geometric mean.

The second application of the strong additivity principle yields

$$D_1(\Pr(X_i, X_k)) = D_1\left(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k\right) D_1\left(X_i \mid X_i \xrightarrow{\Pi_{(i+1,k)}} X_k\right).$$

Since the starting probabilities $\Pr(X_i)$ in the collective diversity $D_1\left(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k\right)$ are those resulting form the random walk $X_0 \xrightarrow{\Pi_{(1,i)}} X_i$, we also have that $D_1\left(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k\right) = D_1\left(X_0 \xrightarrow{\Pi} X_k\right)$.

This gives the desired result

$$D_1\left(X_0 \xrightarrow{\Pi} X_k\right) D_1\left(X_i \mid X_i \xrightarrow{\Pi_{(i+1,k)}} X_k\right) = D_1\left(X_0 \xrightarrow{\Pi_{(1,i)}} X_i\right) D_1\left(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k \mid X_0 \xrightarrow{\Pi_{(1,i)}} X_i\right),$$

from where it follows that

$$D_1\left(X_0 \xrightarrow{\Pi} X_k\right) \leq D_1\left(X_0 \xrightarrow{\Pi_{(1,i)}} X_i\right) D_1\left(X_i \xrightarrow{\Pi_{(i+1,k)}} X_k \mid X_0 \xrightarrow{\Pi_{(1,i)}} X_i\right)$$

if the mean backward diversity is not equal to 1. ■

4.3. Relative diversity

Once the notions of collective and individual diversities have been identified, it is natural to consider the diversity of an individual vertex relative to the collective diversity.

Definition 22 (Relative individual diversity) *The* relative individual V_{end} diversity of $v_0 \in V_{\text{start}}$ with respect to V_{start} along Π *is the relative true diversity between the distribution resulting from a random walk starting at $v_0 \in V_{start}$ (giving its individual diversity), and the distribution resulting from the unconditional random walk starting at random in V_{start} (giving the collective diversity). We denote it by $D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0 = v_0 \,\|\, X_0 \xrightarrow{\Pi} X_k\right)$ and compute it as*

$$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0 = v_0 \,\|\, X_0 \xrightarrow{\Pi} X_k\right) = D_\alpha\left(p_{\Pi|v_0} \,\|\, p_\Pi\right).$$

Using relative true diversities from Section 2.4, other different relative network diversities can computed. Let us consider for example two different meta paths Π_1 and Π_2, such that $V_{\text{start}} = V_{\text{src}}(\Pi_1) = V_{\text{src}}(\Pi_2)$, and $V_{\text{end}} = V_{\text{dst}}(\Pi_1) = V_{\text{dst}}(\Pi_2)$. One diversity measure of interest (to be illustrated in Section 5), is the relative true diversity between distributions on V_{end} resulting from following different meta paths:

$$D_\alpha\left(X_0 \xrightarrow{\Pi_1} X_k \,\|\, X_0 \xrightarrow{\Pi_2} X_k\right) = D_\alpha\left(p_{\Pi_1} \,\|\, p_{\Pi_2}\right).$$

Similarly, though not developed in this article, these computations could be extended to the relative mean individual, and backward diversities.

4.4. Projected diversity

Using the projected edges E_Π for a meta path Π (defined in Section 3.3) we can also define the diversity of the distribution on V_{end} of a constrained random walk starting at $v_0 \in V_{\text{start}}$ and using the edges in E_Π.

Definition 23 (Projected diversity) *Let E_Π be the set of projected edges for meta path* Π. *The* projected V_{end} diversity of $v_0 \in V_{\text{start}}$ along Π *is the true diversity of the distribution on the ending vertices on V_{end} of a random walk starting at $v_0 \in V_{start}$, using the edges in E_Π. We denote it by $D_\alpha\left(X_0 \xrightarrow{E_\Pi} X_k \mid X_0 = v_0\right)$ and compute it as*

$$D_\alpha\left(X_0 \xrightarrow{E_\Pi} X_k \mid X_0 = v_0\right) = D_\alpha\left(p_{E_\Pi|v_0}\right).$$

Note that in the previous definition, $p_{E_\Pi|v_0}$ is the probability distribution from Definition 12 when the meta path is made only of E_Π. Figure 6 illustrates the comparison between individual and projected diversity for two cases using the Shannon diversity. One of these cases results in a projected diversity that is lower than individual diversity, while the other results in a projected diversity that is higher than individual diversity.

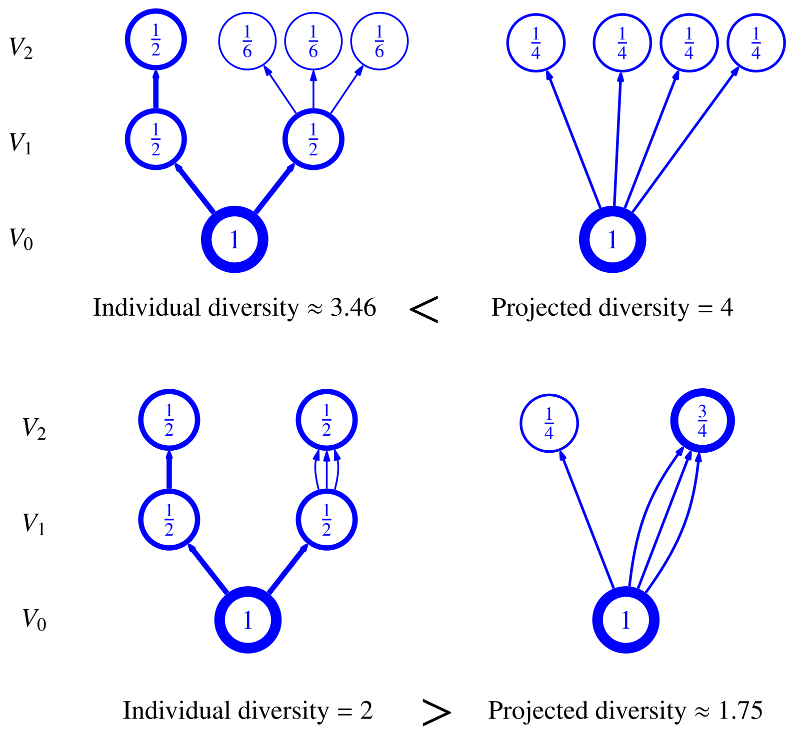


Figure 6: Individual Shannon diversities of a meta path in two heterogeneous information networks, compared with the resulting projected diversities.

Table 2: Summary of the defined diversities along a meta path Π, with $X_0 \in V_{\mathrm{src}}(\Pi)$ and $X_k \in V_{\mathrm{dst}}(\Pi)$.

Diversity	Notation	Expression
Collective	$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k\right)$	$D_\alpha(p_\Pi)$
Individual	$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0 = v_0\right)$	$D_\alpha(p_{\Pi\mid v_0})$
Mean individual	$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0\right)$	$\prod_{v_0 \in V_0} D_\alpha\left(p_{\Pi\mid v_0}\right)^{\Pr(X_0=v_0)}$
Relative individual	$D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0 = v_0 \parallel X_0 \xrightarrow{\Pi} X_k\right)$	$D_\alpha\left(p_{\Pi\mid v_0} \parallel p_\Pi\right)$
Backward individual	$D_\alpha\left(X_0 \mid X_0 \xrightarrow{\Pi} X_k = v_k\right)$	$D_\alpha\left(p_{\Pi^\intercal \mid v_k}\right)$
Projected individual	$D_\alpha\left(X_0 \xrightarrow{E_\Pi} X_k \mid X_0 = v_0\right)$	$D_\alpha\left(p_{E_\Pi \mid v_0}\right)$

4.5. *Summary of network diversities measures*

In this section, we have used the definitions inside the proposed formalism for heterogeneous information networks, in particular that of meta path constrained random walk, to propose different network diversity measures. These include the collective, individual, backward, relative, and projected diversities along a meta path. For each one we have proposed a notation and we have defined their computation using the definitions established in Section 3. Table 2 summarizes the notations and computations of each of the proposed network diversity measures.

In the next section, we will present different application domains for which modeling using heterogeneous information networks is useful. We will show that some quantitative measures computed in different applications are closely related to the network diversity measures, and that their use allows for the consideration of other useful quantitative observables in the modeled systems.

5. Applications

The proposed network diversity measures find numerous applications in many domains where diversity is a relevant information. Information retrieval, and in particular algorithmic recommendation, is one of the areas where they find the most direct application and that best serves as an illustration of their applicability in general. We first illustrate the use of network diversity measures by means of a simple example from recommender systems in Section 5.1. This first example introduces the notations to be used in this section, and the approach chosen to illustrate the application of the network diversity measures. This approach consist in considering a particular heterogeneous information network providing an ontology for data in each domain of application, and showing its network schema (cf. Definition 7). Then, for each application case on each domain, we list several concepts of interest that could be relevant for research questions, together with the explicit expression of the corresponding network diversity measures. These concepts are referred to previous work where they find pertinence when possible. We highlights how these proposed measures can address existing research questions and practices in different research areas, and how they allow for the positing of new ones.

After having introduced a first –simpler– example, we provide a more detailed application case in a recommender systems setting in Section 5.2, explaining the relation between the network diversity measures and several existing existing practices and concepts while also highlighting possible new uses. We then illustrate the use of network diversity measures in the analysis of social networks and news media in Section 5.3. Finally, we provide other examples of applications in ecology, antitrust regulation, and scientometrics in Section 5.4.

5.1. *A simple example*

Let us consider an example with three vertex types: users, items, and types of items. Diverging from the notation established in Section 3.1, for the sake of readability, let us denote these vertex types respectively by V_{users}, V_{items}, and V_{types}. An example of entities represented by items are films, and for types are film genres (e.g. comedy, thriller).

Let us now consider three edge types, indicating items chosen by users, items recommended to users, and classification of items into types. As before, slightly diverging from previously established notation, we denote these edge types as E_{chosen}, $E_{\text{recommended}}$, and E_{types} respectively. Figure 7 illustrates the network schema of the described heterogeneous information network.

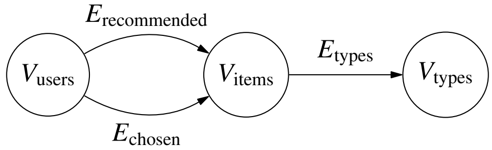


Figure 7: Network schema of a simple heterogeneous information network, where users in V_{users} have chosen and have been recommended items in V_{items}, which are classified into types in V_{types}.

In order to consider random walks constrained to meta paths in this network, let us denote by the capital letter X random vertices in vertex types. Thus, for example, X_{users} is a random vertex in V_{users}. This allows for the consideration of the random walks such as

$$X_{\text{users}} \xrightarrow{E_{\text{chosen}}} X_{\text{items}} \xrightarrow{E_{\text{types}}} X_{\text{types}},$$

for some starting probability distribution $\Pr(X_{\text{users}})$, and constrained to the meta path $\Pi = (E_{\text{chosen}}, E_{\text{types}})$. Throughout this section, we denote a random walk by explicitly writing the vertex types and edge types, as in Definition 11, and for the sake of readability, rather than by its shorter notation $X_{\text{users}} \xrightarrow{\Pi} X_{\text{types}}$.

Using this notation, we identify some concepts of interest related to diversity, and their corresponding network diversity measures. Indeed, we might take interest in the collective type diversity of items recommended to users (cf. Definition 15),

$$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{recommended}}} X_{\text{items}} \xrightarrow{E_{\text{types}}} X_{\text{types}}\right),$$

or in the mean individual type diversity of items recommended to users (cf. Definition 17)

$$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{recommended}}} X_{\text{items}} \xrightarrow{E_{\text{types}}} X_{\text{types}} \mid X_{\text{users}}\right).$$

The network diversity measures allow for the comparison, for example, of the collective type diversity of items recommended and chosen by users, using the relative diversity (cf. Definition 22):

$$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{recommended}}} X_{\text{items}} \xrightarrow{E_{\text{types}}} X_{\text{types}} \,\|\, X_{\text{users}} \xrightarrow{E_{\text{chosen}}} X_{\text{items}} \xrightarrow{E_{\text{types}}} X_{\text{types}}\right).$$

The use of transpose edge types (cf. Definition 18) allows for the referentiation and computation of more complex concepts, such as

$$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{recommended}}} X_{\text{items}} \xrightarrow{E^{\intercal}_{\text{chosen}}} X'_{\text{users}} \xrightarrow{E_{\text{chosen}}} X'_{\text{items}} \xrightarrow{E_{\text{types}}} X_{\text{types}} \mid X_{\text{users}} = u\right),$$

which would otherwise be referred to as the *individual type diversity of the items chosen by users that chose the items recommended to user $u \in V_{users}$*. Some random variables are marked with an apostrophe, e.g. X'_{users}, to indicate that, while they have the same support as the unmarked ones ($\text{supp}(X'_{\text{users}}) = \text{supp}(X_{\text{users}}) = V_{\text{users}}$), they are not the same variable. This is needed in meta paths that include the same vertex type two or more times.

The following examples of application follow this approach: to identify, referentiate, and provide computable expressions for concepts from different domains of research interested in both diversity measures and network representations.

5.2. Recommender Systems

Diversity and diversification of algorithmic recommendations has become one of the leading topics of the recommender systems research community [105, 106]. Through a variety of means, today users have access to a large number of items (e.g. products and services in e-commerce, messages and posts in social media, or news articles in aggregators). While users enjoy an ever-growing offer, it can also become unmanageable for them to consider enough items, or to effectively explore this offer. Recommender systems, developed as early as in 1980s [108], help solving this problem by filtering all the possible items down to a recommended set tailored for each user or group. One recent advance in this field is the recognition of the importance, and the introduction of diversity in recommendations [109, 110].

In recommender systems, diversity can help improving the appreciation of the recommendation quality on the part of users [111, 112]. It also has other applications, such as detecting changes in consumption behavior for context-aware recommenders [114]. As a property of recommendations, diversity has been traditionally captured by a set of related indicators proposed on intuitive bases, called *serendipity*, *discovery*, *novelty*, *dissimilarity* (see Section 8.3 of [115], or [111, 112] for a discussion on the terminology and definitions). These indicators are often computed using past collective choices of items on the part of users [106], or classifications of items into types [20]. To this date, no general framework exists to account for all the proposed diversity indices in recommender systems, nor alternatives on how to exploit richer meta-data structure as those encodable by heterogeneous information networks. This is where the proposed network diversity measures find valuable applications. They accommodate some of the existing concepts from the literature, they extend the measurement of diversity to more complex data structures that can include meta-data on users and items, and they give formal explicit expressions to computable quantities related to new and existing research questions in this field.

For the purposes of illustrations let us consider a heterogeneous information network giving an ontology to complex data related to a situation in which we have recommended different types of items to users. Figure 8 shows the network schema of the heterogeneous information network to be considered in this example. Let us consider the following vertex types for the example (as before abusing notation, identifying types of vertices or edges with the actual subsets of vertices):

- A vertex type of users V_U;
- Two vertex types for items: V_{I_1} (e.g. films) and V_{I_2} (e.g. series);
- Two vertex types for item classification: V_{T_1} (e.g. channels/distributors) and V_{T_2} (e.g. genre);
- Two vertex types of user groups: V_{G_1} (e.g. demographic group) and V_{G_2} (e.g. location).

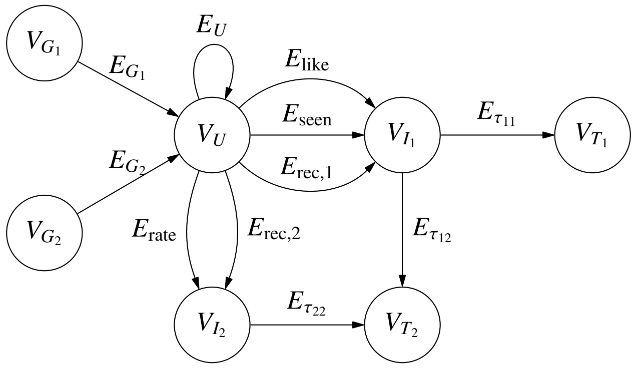


Figure 8: Network schema of a heterogeneous information network in a setting from recommender systems, where users in V_U, belonging to groups V_{G_1} and V_{G_2} interact and are recommended two different sets of items V_{I_1} and V_{I_2}, which are classified using types in V_{T_1} and V_{T_2}.

In order to consider random walks constrained to meta path, and following the example from Section 5.1, we denote with capital letter X the random variables supported by a vertex type. For example, X_U is a random vertex in V_U and X_{I_1} is a random vertex in V_{I_1}.

In the heterogeneous information network illustrated in Figure8 we also consider different edge types:

- An edge type of edges between users E_U (e.g. users that follow other users in a social network);
- Edge types from groups to the users: E_{G_1} (e.g. associating users with demographic groups) and E_{G_2} (e.g. associating users with locations);
- Edge types indicating classification of items into types: $E_{\tau_{11}}$ and $E_{\tau_{12}}$ for V_{I_1}, and $E_{\tau_{22}}$ for V_{I_2};
- Vertex types for representing when users have liked (E_{like}), seen (E_{seen}), or rated (E_{rate}) an item, or for representing when users have been recommended items ($E_{\text{rec},1}$ and $E_{\text{rec},2}$).

All of the elements in the proposed example are useful for representing common practices in recommender systems. Settings for recommendation where there are two –or more– types of items (V_{I_1} and V_{I_2} in our example) are common in cross-domain recommendation [116], and more generally in heterogeneous information network recommendation [27]. Relations between users and items can be of different kinds in recommendation settings: edges can be used to indicate that a user has rated an item in *explicit feedback* –or scoring, or noting– systems (E_{rate} in the example), or to indicate that a user has liked an item in *implicit feedback* systems (E_{like} in the example). Some recommender systems and diversity measures can take into account whether a user has previously seen an item [107] (E_{seen} in the example). Also, settings where meta-data are associated with users is very common in demographic, or location filtering [117], and are represented in the example by using vertex types V_{G_1} and V_{G_2}. Finally, edges between users signaling relations such as friendship of a user *following* another one on social networks (E_U in the example) can also be exploited in recommendations [118, 119], and certainly in diversity computations. As stated before, Figure 8 represents the network schema (cf. Definition 7) or our example.

Most diversity computations in recommender systems are concerned with providing a measure of diversity for the items recommended to a user, or an aggregation of this quantity for all users. Diversity between items can be computed, for example, with respect to a classification of the items [120] (e.g. genres for films) . In the proposed framework, this concept can be captured by the individual diversity. Let us imagine that V_U are users, that V_{I_2} are films, and that V_{T_2} are film genres (e.g. comedy, thriller, etc.). The individual genre (V_{T_2}) diversity of films (V_{I_2}) recommended to a user $u \in V_U$ is

$$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},2}} X_{I_2} \xrightarrow{E_{\tau_{12}}} X_{T_1} \mid X_U = u\right).$$

Similarly, the mean genre (V_{T_2}) diversity of films (V_{I_2}) recommended to all users V_U is the mean individual diversity

$$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},2}} X_{I_2} \xrightarrow{E_{\tau_{12}}} X_{T_1} \mid X_U\right),$$

which can be computed as a geometric mean by choosing $X_U \sim \text{Uniform}(V_U)$ for the starting of the meta path-constrained random walk.

In another classic setting, the diversity of an item is computed according to the number of users that have previously chosen or liked it (sometimes called *novelty* [121]). In the proposed framework, an aggregation of this quantity for items proposed to all users corresponds to the following network diversity measure:

$$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E^{\intercal}_{\text{like}}} X'_U \mid X_U\right).$$

More interestingly, other relevant quantities expressible as network diversity measures have no explicit expression in other frameworks of the literature. The clearest example is the comparison between the mean individual and collective recommended diversities: for example, $D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \mid X_U\right)$ versus $D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1}\right)$. Distinguishing these two concepts is important when taking interest in diversity beyond its use as a quality of recommendations; for example, when studying phenomena such as filter bubbles (cf. Figure 5).

Let us present in a schematic fashion, in Table 3, different examples of concepts related to diversity that are of interest for research questions in the domain of recommender system, together with the respective quantities that can be identified, expressed, and computed as network diversity measures.

5.3. Social media studies, echo chambers, and filter bubbles

The study of social media has been developed into a large and ever growing wealth of results. The importance of studies about the creation, transmission, and consumption of information in social networks has become crucial. Heterogeneous information networks is a natural formalism for the treatment of these objects, as they can accommodate a variety of entities (e.g. posts, accounts, media outlets, tags, keywords) interacting through many different relations (e.g. users publishing posts, mentioning other users, following between users, using tags in publication). More complex and abstract data is often analyzed in these studies, such as political affiliations for users and media outlets [122, 123]. The analysis of phenomena such as echo chambers and filter bubbles through the measurement of diversity of information consumption is an established practice [21, 22, 124, 125, 23]. The settings of different social media studies vary, but concrete example are the study of the *Leave* and *Remain* Brexit campaigns on Twitter [126], or the exchange of information between US Democrats and Republicans on Facebook [127].

In this section, we illustrate the use of network diversity measures in the study of information exchange in social networks. In order to do so, we consider, as before, a heterogeneous information network for illustration, now in the

Table 3: Schematic representation of examples of concepts related to diversity in recommender systems and the network diversity measures that can be used to address them in quantitative studies.

Examples of concepts expressible in research questions	Corresponding network diversity measure
Comparison between mean individual, and collective diversity of recommendation of items V_{I_1} according to types V_{T_1}	$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \parallel X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \mid X_U\right)$
Comparison between collective diversity of recommended and liked items V_{I_1} according types V_{T_1}	$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \parallel X_U \xrightarrow{E_{\text{like}}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1}\right)$
Comparison between collective diversity of recommendations of items in V_{I_1} (e.g. films) and V_{I_2} (e.g. series) according to types V_{T_2} (e.g. genres)	$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_2} \parallel X_U \xrightarrow{E_{\text{rec},2}} X_{I_2} \xrightarrow{E_{\tau_{22}}} X_{T_2}\right)$
Diversity of items V_{I_1} recommended to friends of $u \in V_U$ according to types V_{T_1}	$D_\alpha\left(X_U \xrightarrow{E_U} X'_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_1} \mid X_U = u\right)$
Diversity of items V_{I_1} liked by group $g \in V_{G_1}$ according to types V_{T_1}	$D_\alpha\left(X_{G_1} \xrightarrow{E_{G_1}} X_U \xrightarrow{E_{\text{like}}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_1} \mid X_{G_1} = g\right)$
Diversity of users V_U that liked items V_{I_1} of type $t \in V_{T_2}$	$D_\alpha\left(X_U \mid X_U \xrightarrow{E_{\text{like}}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_2} = t\right)$
Diversity of types V_{T_2} chosen by user $u \in V_U$ through their choices of items in V_{I_1}	$D_\alpha\left(X_U \xrightarrow{E_{(E_{\text{like}}, E_{\tau_{12}})}} X_{T_2} \mid X_U = u\right)$

setting of data traces of activity in social networks and media. Figure 9 shows a network schema chosen to illustrate the applicability of the proposed network diversity measures to this context. The heterogeneous information network schema of Figure 9 considers: users that post or share posts (or *tweets*, or blog entries, or comments in forums), users that can follow (or befriend) other users, posts that can mention users, include tags (e.g. *hashtags*), include topics (detectable, for example, by matching strings or using topic discovery methods), or even link to articles through an URL address. In many settings of interest articles can be associated with media outlets, which can in turn be identified with groups or affiliations (e.g. political parties).

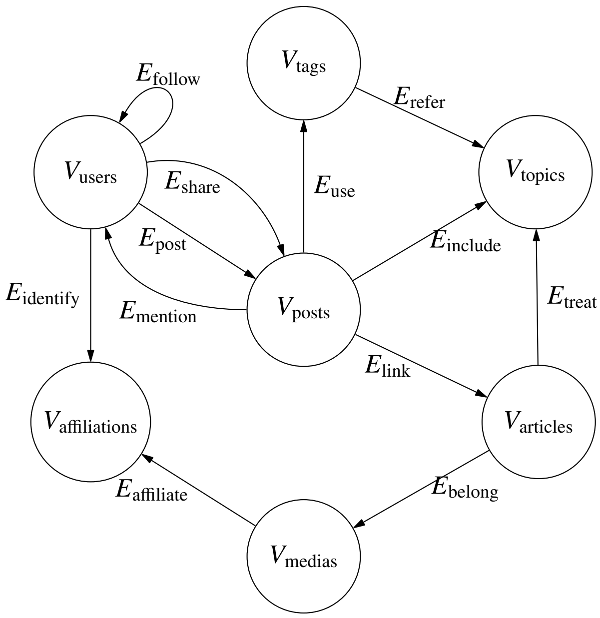


Figure 9: Network schema of a heterogeneous information network in a setting from social networks and media studies, suited for the study of echo chambers and filter bubbles.

The elements chosen for the illustration can accommodate different aspects considered in new media studies. For example [122] considers relations of political affiliation of users and interactions between them, and analyzes the notion of diversity of Twitter posts according to the political communities they have reached. Other studies also consider the use of entropy measures over distributions representing the proportion of users that browse given information sources [5]. Some studies, for example [128], explicitly consider networks of information items (e.g. blog posts) and the concepts that they use.

As with the example of recommender systems settings, we present in a schematic fashion (in Table 4) different concepts related to diversity that are of interest for research questions the field of social networks and media studies, together with quantities that can be computed as network diversity measures.

Table 4: Schematic representation of examples of concepts related to diversity in social networks and media studies, and the corresponding network diversity measures that can be used to address them in quantitative studies.

Examples of concepts expressible in research questions	Corresponding network diversity measure
Collective diversity of affiliations of users	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}}\right)$
Collective affiliation diversity of users through the contents they share	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{share}}} X_{\text{posts}} \xrightarrow{E_{\text{link}}} X_{\text{articles}} \xrightarrow{E_{\text{belong}}} X_{\text{medias}} \xrightarrow{E_{\text{affiliate}}} X_{\text{affiliations}}\right)$
Individual topic diversity of user $u \in V_{\text{users}}$ through posting	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{include}}} X_{\text{topics}} \mid X_{\text{users}} = u\right)$
Individual topic diversity of user $u \in V_{\text{users}}$ through posting of followers	$D_\alpha\left(X_{\text{users}} \xrightarrow{E^{\intercal}_{\text{follow}}} X'_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{include}}} X_{\text{topics}} \mid X_{\text{users}} = u\right)$
Individual affiliation diversity of users mentioned by user $u \in V_{\text{users}}$	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{mention}}} X'_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}} \mid X_{\text{users}} = u\right)$
Diversity of affiliation groups that treat topic $t \in V_{\text{topics}}$ in articles	$D_\alpha\left(X_{\text{affiliations}} \mid X_{\text{affiliations}} \xrightarrow{E^{\intercal}_{\text{affiliate}}} X_{\text{medias}} \xrightarrow{E^{\intercal}_{\text{belong}}} X_{\text{articles}} \xrightarrow{E_{\text{treat}}} V_{\text{topics}} = t\right)$
Comparison of affiliation diversity of users according to what they posts and to their political identification	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{mention}}} X'_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}} \parallel X_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}}\right)$

5.4. Other examples

Many other domains of studies are concerned by both diversity measures and network models. In the rest of this section, we provide illustrations of how the diversity measures could find applications in identifying, referencing, and providing computable expressions for relevant concepts related to diversity in ecology, competition law, and scientometrics.

Ecology

Ecology is a research domain concerned by the concept of diversity, as identified and commented in Section 2.2. Many of the advancements in diversity measures come from this community (e.g. [13]). One prominent concept in this domain is that of the diversity of species in an habitat. For the computation of quantitative indices of this diversity, individuals from different species are counted, or their number is estimated. From their apportionment into the species present in an habitat, a diversity is computed and reported.

But ecology, interested too in interactions (among organisms), is growingly concerned by graph representations and models. One of such interactions, also related to diversity, is represented by the so-called *food webs* [129]: networks models that describe species that feed on other species. In the past, there have been efforts to use graph formalisms to treat food webs [130, 131]. Similarly, other relations between species have been described using graphs, such as relations of parasitation [132]. Another subject of interest in ecology is the description of habitats and their connectedness, for which there have also been approaches using graph theory to describe these connections [133, 134].

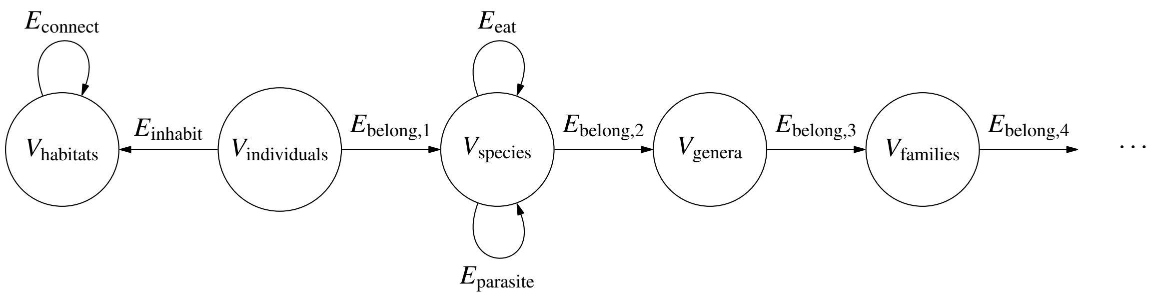
All of the cited elements can be treated using heterogeneous information networks. And using network diversity measures, different concepts of interest related to diversity can be computed over them. Let us consider for example a heterogeneous information network with vertex types for habitats (V_{habitats}), for individuals ($V_{\text{individuals}}$), for species (V_{species}), for genera (V_{genera}), for families (V_{families}), and so on as needed according to the study that this example might interest.

Let as also consider for our example edges grouped in edge types. Edge type E_{connect} is that of edges between habitats, indicating wether an individual can access a given habitat from another one. Edge type E_{inhabit} is used to represent which individuals inhabit which habitats. Edge types E_{eat} and E_{parasite} are used to represent relations between species; which species eat which species, and similarly for parasitation. Edge type $E_{\text{belong},1}$ is used to represent which individual belong which species. Finally, edge types $E_{\text{belong},2}$ and $E_{\text{belong},3}$ contain edges indicating which species belong to which genus, and which genus belong to which family (species, genus, and family are three of the eight major taxonomic ranks in biological classification).

This setting can accomodate the common practice of the measurement of the –bio– diversity of a habitat $h \in V_{\text{habitats}}$, which in network diversity measures finds the expression

$$D_\alpha\left(X_{\text{habitats}} \xrightarrow{E^\intercal_{\text{inhabit}}} X_{\text{individuals}} \xrightarrow{E_{\text{belong},1}} X_{\text{species}} \mid X_{\text{habitats}} = h\right),$$

with $\alpha = 0$ giving the richness biodiversity, and $\alpha = \infty$ the Berger-Parker biodiversity of the habitat. Figure 10 illustrates the network schema of the described heterogeneous information network, and a table with some examples of diversity-related concepts and their expressions as network diversity measures.



Examples of concepts expressible in research questions	Corresponding network diversity measure
Species diversity in habitat $h \in V_{\text{habitats}}$	$D_\alpha\left(X_{\text{species}} \mid X_{\text{species}} \xrightarrow{E^\intercal_{\text{belong},1}} X_{\text{individuals}} \xrightarrow{E_{\text{inhabit}}} X_{\text{habitats}} = h\right)$
Genera diversity in habitat $h \in V_{\text{habitats}}$	$D_\alpha\left(X_{\text{genera}} \mid X_{\text{genera}} \xrightarrow{E^\intercal_{\text{belong},2}} X_{\text{species}} \xrightarrow{E^\intercal_{\text{belong},1}} X_{\text{individuals}} \xrightarrow{E_{\text{inhabit}}} X_{\text{habitats}} = h\right)$
Species diversity of habitats adjacent to those where a species $s \in V_{\text{species}}$ is present	$D_\alpha\left(X_{\text{species}} \xrightarrow{E^\intercal_{\text{belong},1}} X_{\text{individuals}} \xrightarrow{E_{\text{inhabit}}} X_{\text{habitats}} \xrightarrow{E_{\text{connect}}} X'_{\text{habitats}} \xrightarrow{E^\intercal_{\text{inhabit}}} X'_{\text{individuals}} \xrightarrow{E_{\text{belong},1}} X'_{\text{species}} \mid X_{\text{species}} = s\right)$
Species diversity of the predators of species that parasite a species $s \in V_{\text{species}}$	$D_\alpha\left(X_{\text{species}} \mid X_{\text{species}} \xrightarrow{E_{\text{eat}}} X'_{\text{species}} \xrightarrow{E_{\text{parasite}}} X''_{\text{species}} = s\right)$
Comparison of diversity in habitats $h_1, h_2 \in V_{\text{habitats}}$	$D_\alpha\left(X_{\text{habitats}} \xrightarrow{E^\intercal_{\text{inhabit}}} X_{\text{individuals}} \xrightarrow{E_{\text{belong},1}} X_{\text{species}} \mid X_{\text{habitats}} = h_1 \parallel X'_{\text{habitats}} \xrightarrow{E^\intercal_{\text{inhabit}}} X'_{\text{individuals}} \xrightarrow{E_{\text{belong},1}} X'_{\text{species}} \mid X'_{\text{habitats}} = h_2\right)$

Figure 10: Network schema of an example from ecology, and table with examples of diversity-related concepts and their expression as network diversity measures.

Competition law

Many developments and applications of concentration measures are found in the community of economics, antitrust regulation, and competition law. As it was shown in Section 2.2, concentration is a concept for which indices are the reciprocal of those for the concept of diversity. In this domain, concentration or diversity indices are used to measure the degree to which some firms concentrate the production of units (or the provision of services) in an industry. Let us consider, for example, the classification and apportionment of the tons of steel produced –in a given period in a country– into the firms that produced them. From this apportionment or distribution, the concentration of the steel industry can be measured quantitatively with diversity measures. This is the subject of the doctoral thesis of O. C. Herfindahl, for which he developed what is now known as the Herfindahl-Hirschman Index [68]. The quantitative measurement of the concentration of an industry allows for important comparisons made by industry regulators, such as, for example, the degree of concentration of an industry should a given merger or an acquisition be allowed.

This exercise in measurement of industrial concentration, and the detection and limitation of monopolistic behavior, is made significantly more difficult by the existence of cross-ownership, or cross-control relation between firms.

Cross-ownership refers to situations in which firms from an industry are mutually owned among them in complex, network-like relations (in a simple example between two firms A and B, firm A owns a part of firm B, and firm B owns a part of firm A). Cross-control refers, similarly, to situations where firms control actions of other firms in the same industry producing complex relations of control in a network-like fashion.

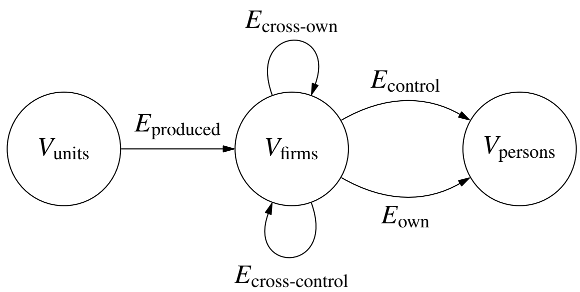
Specialized economic literature accounts for many examples of cases that challenge the application of the aforementioned procedure in the regulation [135, 136], and that address the complex structure of co-ownership networks and their importance in regulation [137, 138]. This makes graph-theoretical approaches good candidates for making advances in the measurement of concentration in industries [139]. In particular, the proposed network diversity measures provide tools that allow to measure many concepts of interest in antitrust regulation and competition law when dealing with network structures.

To illustrate this, let us consider an example consisting of a heterogeneous information network with three vertex types: that of vertices representing produced units of services V_{units} (e.g. tons of steel, barrels of oil, or clients of portable phone services), that of vertices representing firms V_{firms} that produce those units or provide those services, and that for persons V_{persons} that own the firms. To model the relations between these entities represented by vertex types, let us consider five edge types: E_{produced}, linking each unit to the firm that produced it, E_{own}, linking firms with the persons that own them, $E_{\text{cross-own}}$, linking firms between them by relations of cross-ownership, and similarly, E_{control} and $E_{\text{cross-control}}$ linking firms between them by relations of control and cross-control (for example, having the right to choose a member of the board on a firm). For type edges E_{own}, $E_{\text{cross-own}}$, E_{control}, and $E_{\text{cross-control}}$, the multiplicity of edges can account for the units on which property or control is represented, such as, for example, stocks or members of the boards of the firms. For example, if ownership of a firm is represented by 10 stocks, it will have 10 edges, that can belong to edges types E_{own} or $E_{\text{cross-own}}$.

In this setting, the common measurement of industry diversity, is expressed as

$$D_\alpha\left(X_{\text{units}} \xrightarrow{E_{\text{produced}}} X_{\text{firms}}\right),$$

which becomes the Herfindahl-Hirschman Diversity (reciprocal of the Herfindahl-Hirschman Index) by choosing $\alpha = 2$. Figure 11 illustrates the heterogeneous information network described for this example, with a table of examples of concepts related to diversity and expressible using the proposed network diversity measures.



Examples of concepts expressible in research questions	Corresponding network diversity measure
Industry diversity with cross-ownership relations	$D_\alpha\left(X_{\text{units}} \xrightarrow{E_{\text{produced}}} X_{\text{firms}} \xrightarrow{E_{\text{cross-own}}} X'_{\text{firms}}\right)$
Industry diversity according to persons with cross-ownership relations	$D_\alpha\left(X_{\text{units}} \xrightarrow{E_{\text{produced}}} X_{\text{firms}} \xrightarrow{E_{\text{cross-own}}} X'_{\text{firms}} \xrightarrow{E_{\text{own}}} X_{\text{persons}}\right)$
Diversity of cross-ownership of a firm $f \in V_{\text{firms}}$	$D_\alpha\left(X_{\text{firms}} \xrightarrow{E_{\text{cross-own}}} X'_{\text{firms}} \mid X_{\text{firms}} = f\right)$
Comparison of diversity of ownership and control of a firms by persons	$D_\alpha\left(X_{\text{firms}} \xrightarrow{E_{\text{own}}} X_{\text{persons}} \parallel X'_{\text{firms}} \xrightarrow{E_{\text{control}}} X'_{\text{persons}}\right)$

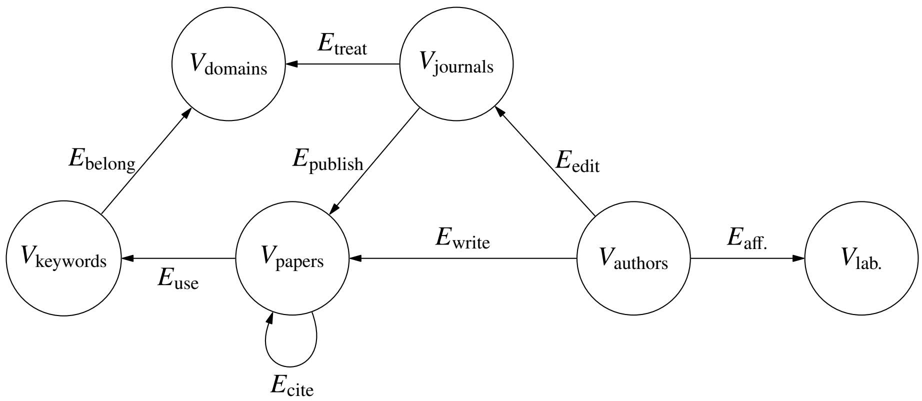
Figure 11: Network schema of a cross-ownership and cross-control heterogeneous information network in a setting from *antitrust* regulation or competition law, where products are apportioned in the firms that produced them, which can be cross-owned or cross-controlled among them.

Scientometrics

Scientometrics, within the field of bibliometrics, is concerned by the measurement and analysis of scientific literature. Overlap with information systems, scientometrics study, for example, the *importance* of publications in networks of citations using metrics such as the *Impact Factor*, or the *Science Citation Index*. In networks including other entities such as authors, other measurements include the *h-index*, an index for the productivity and citation impact of scholars. Recent studies have used heterogeneous information networks to represent data including other entities, such as journals and conferences in order to extract extended measuresments [140].

The study of networks modeling and representing scientific publication is of interest for other reasons too. The diversity of topics explored by scientific communities is a concept of interest, for example, in public policy [141], an in the general in the understanding and description of the structure of scientific communities [143, 144]. Other practical applications for the measurement of diversities in citation networks is the maintenance of classification systems [142].

In this section we illustrate the way in which the proposed network diversity measures can address some of the concepts relevant to these areas of research by means of an example. Let us consider a heterogeneous information network in a setting of citation networks in scientific literature, consisting of the following vertex types: authors V_{authors}, laboratories $V_{\text{lab.}}$ (or affiliation institutions), journals V_{journals}, scientific articles V_{papers}, keywords used by these articles V_{keywords}, and domains of research V_{domains} (e.g. ecology, economics). We also consider edge types to represent relations between these entities (see Figure 12): affiliation $E_{\text{aff.}}$ of authors to institutions, edition (or peer-review) E_{edit} of journals by authors, writing E_{write} of articles by authors, the use of keywords E_{use} in articles, the association of keywords E_{belong} with domains, the publishing E_{publish} of articles by journals, and the declared treatment E_{treat} of research domains by journals. Figure 12 illustrates the heterogeneous information network described for this example, with a table of examples of concepts related to diversity and expressible using the proposed network diversity measures.



Examples of concepts expressible in research questions	Corresponding network diversity measure
Diversity of keywords used by author $a \in V_{\text{authors}}$	$D_\alpha\left(X_{\text{authors}} \xrightarrow{E_{\text{write}}} X_{\text{papers}} \xrightarrow{E_{\text{use}}} X_{\text{keywords}} \mid X_{\text{authors}} = a\right)$
Comparison of diversity of domains addressed in publications and in editing (or peer-reviewing) by author $a \in V_{\text{authors}}$	$D_\alpha\left(X_{\text{authors}} \xrightarrow{E_{\text{write}}} X_{\text{papers}} \xrightarrow{E_{\text{use}}} X_{\text{keywords}} \xrightarrow{E_{\text{belong}}} X_{\text{domains}} \mid X_{\text{authors}} = a \parallel X'_{\text{authors}} \xrightarrow{E_{\text{edit}}} X_{\text{journals}} \xrightarrow{E_{\text{treat}}} X'_{\text{domains}} \mid X'_{\text{authors}} = a\right)$
Diversity of domains addressed by citations by authors of laboratory $l \in V_{\text{lab.}}$	$D_\alpha\left(X_{\text{lab.}} \xrightarrow{E_{\text{aff.}}^{\intercal}} X_{\text{authors}} \xrightarrow{E_{\text{write}}} X_{\text{papers}} \xrightarrow{E_{\text{cite}}} X'_{\text{papers}} \xrightarrow{E_{\text{use}}} X_{\text{keywords}} \xrightarrow{E_{\text{belong}}} X_{\text{domains}} \mid X_{\text{lab.}} = l\right)$

Figure 12: Network schema of a heterogeneous information network in a setting from scientometrics, and table of examples of concepts related to diversity and expressible using the proposed network diversity measures.

6. Conclusions

This article presents a formal framework for the measurement of diversity on heterogeneous information networks. This allows for the extension of the application of diversity measures, from classification modeled by apportioning into distributions, to data represented in network structures.

By presenting a succinct theory resulting from the imposition of desirable properties of axioms, we organize diversity measures across a wide spectrum of domains into a family of functions defined by a single parameter: the true diversities. Providing a formalism for heterogeneous information networks and constrained random walks on it, we consider different probability distributions on which diversity measures are computed. These diversity measures are related to the structure of the heterogeneous information network, and thus to the phenomenon or object that it represents. Diversity measures are also related to the different ways in which distributions are computed, which allows us to distinguish several types of diversities: collective, individual, mean individual, backward, relative, and projected diversities. Some of these network diversities relate to existing measures of the literature, that we framed into a comprehensive framework. But they also allow for the treatment of new concepts related to diversity in networks. We provide examples of this through illustrations of applications in several domains.

The main contributions of this article are:

- A proposition of an axiomatic theory of diversity measures that allows us to present most of their uses across several domains with an existing single-parameter family of functions.
- A formalization of the concepts and tools to describe and process heterogeneous information networks, that have been gaining attention in the representation learning, and information retrieval communities (in particular in recommender systems).
- The definition of several *network diversity measures*, resulting from the application of true diversities to the probability distributions computable in the heterogeneous information network formalism. These network diversity measures allow for the referentiation, expression, and computation of concepts relevant to diversity in networks, extending the use of diversity, from systems of classification and apportionment, to systems best described by network-structured data.
- The mapping of some of the network diversity measures to existing quantitative measurements which are widespread in different fields, and the development of new applications through examples in recommender systems, social media studies, ecology, competition law, and scientometrics.

Acknowledgement

This work has been partially funded by the European Commission H2020 FETPROACT 2016-2017 program under grant 732942 (ODYCCEUS) and by the French National Agency of Research (ANR) under grant ANR-15-CE38-0001 (AlgoDiv).

We thank Hong-Lan Botterman and Matthieu Latapy for their comments.

References

[1] K. S. McCann, The diversity–stability debate, Nature 405 (6783) (2000) 228.

[2] P. Geroski, et al., The choice between diversity and scale, Davis, E., Geroski, PA, Kay, J. A., Manning A., Smales, C., Smith, SR and Szymanski, S.(eds.) (1992) 29–45.

[3] J. Aczél, Z. Daróczy, On measures of information and their characterizations, New York (1975) 168.

[4] A. Rényi, et al., On measures of entropy and information, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, The Regents of the University of California, 1961, pp. 547–561.

[5] D. Nikolov, D. F. Oliveira, A. Flammini, F. Menczer, Measuring online social bubbles, PeerJ Computer Science 1 (2015) e38.

[6] J. Kulshrestha, M. B. Zafar, L. E. Noboa, K. P. Gummadi, S. Ghosh, Characterizing information diets of social media users., in: ICWSM, 2015, pp. 218–227.

[7] R. K. Ursem, Diversity-guided evolutionary algorithms, in: International Conference on Parallel Problem Solving from Nature, Springer, 2002, pp. 462–471.

[8] J. Riget, J. S. Vesterstrøm, A diversity-guided particle swarm optimizer-the arpso, Dept. Comput. Sci., Univ. of Aarhus, Aarhus, Denmark, Tech. Rep 2 (2002) 2002.

[9] A. Stirling, A general framework for analysing diversity in science, technology and society, Journal of the Royal Society Interface 4 (15) (2007) 707–719.

[10] C. E. Shannon, A mathematical theory of communication, Bell system technical journal 27 (3) (1948) 379–423.

[11] C. Gini, Measurement of inequality of incomes, The Economic Journal 31 (121) (1921) 124–126.

[12] S. A. Rhoades, The herfindahl-hirschman index, Fed. Res. Bull. 79 (1993) 188.

[13] M. O. Hill, Diversity and evenness: a unifying notation and its consequences, Ecology 54 (2) (1973) 427–432.

[14] D. Encaoua, A. Jacquemin, Degree of monopoly, indices of concentration and threat of entry, International economic review (1980) 87–105.

[15] S. R. Chakravarty, W. Eichhorn, An axiomatic characterization of a generalized index of concentration, Journal of Productivity Analysis 2 (2) (1991) 103–112.
[16] F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information, Harvard University Press, Cambridge, MA, USA, 2015.
[17] E. Pariser, The Filter Bubble: What The Internet Is Hiding From You, Penguin Books Limited, 2011.
[18] A. Datta, J. Makagon, D. Mulligan, M. Tschantz, Discrimination in online personalization: A multidisciplinary inquiry, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, ACM, 2018.
[19] R. Courtland, Bias detectives: the researchers striving to make algorithms fair, Nature 558 (7710) (2018) 357–360.
[20] C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th international conference on World Wide Web, ACM, 2005, pp. 22–32.
[21] B. Smyth, P. McClave, Similarity vs. diversity, in: Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, ICCBR '01, Springer-Verlag, 2001, pp. 347–361.
[22] A.-A. Stoica, C. Riederer, A. Chaintreau, Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity, in: Proceedings of the 2018 World Wide Web Conference, ACM, 2018, pp. 923–932. doi:10.1145/3178876.3186140.
[23] M. Schedl, P. Knees, F. Gouyon, New paths in music recommender systems research, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM, 2017, pp. 392–393.
[24] Y. Sun, J. Han, Mining heterogeneous information networks: a structural analysis approach, Acm Sigkdd Explorations Newsletter 14 (2) (2013) 20–28.
[25] Y. Sun, Y. Yu, J. Han, Ranking-based clustering of heterogeneous information networks with star network schema, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 797–806.
[26] M. Ji, J. Han, M. Danilevsky, Ranking-based classification of heterogeneous information networks, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 1298–1306.
[27] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, J. Han, Personalized entity recommendation: A heterogeneous information network approach, in: Proceedings of the 7th ACM international conference on Web search and data mining, ACM, 2014, pp. 283–292.
[28] T.-y. Fu, W.-C. Lee, Z. Lei, Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, 2017, pp. 1797–1806.
[29] M. Sydow, M. Pikuła, R. Schenkel, The notion of diversity in graphical entity summarisation on semantic knowledge graphs, Journal of Intelligent Information Systems 41 (2) (2013) 109–149.
[30] Y. Chen, C. Wang, Hine: heterogeneous information network embedding, in: International Conference on Database Systems for Advanced Applications, Springer, 2017, pp. 180–195.
[31] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, A. Arenas, Mathematical formulation of multilayer networks, Physical Review X 3 (4) (2013) 041022.
[32] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, A. Arenas, Ranking in interconnected multilayer networks reveals versatile nodes, Nature communications 6 (2015) 6868.
[33] Y. Sun, J. Han, X. Yan, P. S. Yu, T. Wu, Pathsim: Meta path-based top-k similarity search in heterogeneous information networks, Proceedings of the VLDB Endowment 4 (11) (2011) 992–1003.
[34] N. Lao, W. W. Cohen, Relational retrieval using a combination of path-constrained random walks, Machine learning 81 (1) (2010) 53–67.
[35] Y. Xiong, Y. Zhu, S. Y. Philip, Top-k similarity join in heterogeneous information networks, IEEE Transactions on Knowledge and Data Engineering 27 (6) (2014) 1710–1723.
[36] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, J. Han, Recommendation in heterogeneous information networks with implicit user feedback, in: Proceedings of the 7th ACM conference on Recommender systems, ACM, 2013, pp. 347–350.
[37] C. Shi, B. Hu, W. X. Zhao, S. Y. Philip, Heterogeneous information network embedding for recommendation, IEEE Transactions on Knowledge and Data Engineering 31 (2) (2018) 357–370.
[38] K. Yu, S. Yu, V. Tresp, Soft clustering on graphs, in: Advances in neural information processing systems, 2006, pp. 1553–1560.
[39] R.-H. Li, J. X. Yu, Scalable diversified ranking on large graphs, IEEE Transactions on Knowledge and Data Engineering 25 (9) (2012) 2133–2146.
[40] D. Hristova, M. J. Williams, M. Musolesi, P. Panzarasa, C. Mascolo, Measuring urban social diversity using interconnected geo-social networks, in: Proceedings of the 25th international conference on world wide web, International World Wide Web Conferences Steering Committee, 2016, pp. 21–30.
[41] B. King, R. Jha, D. R. Radev, Heterogeneous networks and their applications: Scientometrics, name disambiguation, and topic modeling, Transactions of the Association for Computational Linguistics 2 (2014) 1–14.
[42] S. Nandanwar, A. Moroney, M. N. Murty, Fusing diversity in recommendations in heterogeneous information networks, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM, 2018, pp. 414–422.
[43] F. Vahedian, R. Burke, B. Mobasher, Meta-path selection for extended multi-relational matrix factorization, in: The Twenty-Ninth International Flairs Conference, 2016.
[44] X. Yu, Y. Sun, B. Norick, T. Mao, J. Han, User guided entity similarity search using meta-path selection in heterogeneous information networks, in: Proceedings of the 21st ACM international conference on Information and knowledge management, Acm, 2012, pp. 2025–2029.
[45] E. P. Odum, Fundamentals of ecology, WB Saunders company, 1959.
[46] G. Helfman, B. B. Collette, D. E. Facey, B. W. Bowen, The diversity of fishes: biology, evolution, and ecology, John Wiley & Sons, 2009.
[47] T. M. Bezemer, W. H. Van Der Putten, Ecology: diversity and stability in plant communities, Nature 446 (7135) (2007) E6.
[48] S. J. McNaughton, Diversity and stability of ecological communities: a comment on the role of empiricism in ecology, The American Naturalist 111 (979) (1977) 515–525.
[49] R. M. May, Patterns of species abundance and diversity, Ecology and evolution of communities (1975) 81–120.
[50] J. M. Smith, Trees, bundles or nets?, Trends in ecology & evolution 4 (10) (1989) 302–304.

[51] S. Gillett, In praise of policy diversity, position paper for oii broadband forum.
[52] W. H. Dutton, S. E. Gillett, L. W. McKnight, M. Peltu, Bridging broadband internet divides: reconfiguring access to enhance communicative power, Journal of Information Technology 19 (1) (2004) 28–38.
[53] G. Silverberg, G. Dosi, L. Orsenigo, Innovation, diversity and diffusion: A self-organisation model, The Economic Journal 98 (393) (1988) 1032–1054.
URL `http://www.jstor.org/stable/2233718`
[54] H. Nowotny, P. Scott, M. Gibbons, P. B. Scott, Re-thinking science: Knowledge and the public in an age of uncertainty, Cambridge: Polity Press, 200.
[55] E. V. Shevchenko, D. V. Talapin, N. A. Kotov, S. O'brien, C. B. Murray, Structural diversity in binary nanoparticle superlattices, Nature 439 (7072) (2006) 55.
[56] E. D. Schneider, J. J. Kay, Life as a manifestation of the second law of thermodynamics, Mathematical and computer modelling 19 (6-8) (1994) 25–48.
[57] G. Grabher, D. Stark, Organizing diversity: evolutionary theory, network analysis and postsocialism, Regional studies 31 (5) (1997) 533–544.
[58] W. R. Ashby, Requisite variety and its implications for the control of complex systems, in: Facets of systems science, Springer, 1991, pp. 405–417.
[59] F. J. Dyson, Statistical theory of the energy levels of complex systems. i, Journal of Mathematical Physics 3 (1) (1962) 140–156.
[60] H.-X. Yang, Z.-X. Wu, C. Zhou, T. Zhou, B.-H. Wang, Effects of social diversity on the emergence of global consensus in opinion dynamics, Physical Review E 80 (4) (2009) 046108.
[61] A. Stirling, On the economics and analysis of diversity, Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper 28 (1998) 1–156.
[62] R. H. MacArthur, Patterns of species diversity, Biological reviews 40 (4) (1965) 510–533.
[63] N. J. Gotelli, R. K. Colwell, Estimating species richness, Biological diversity: frontiers in measurement and assessment 12 (2011) 39–54.
[64] C. E. Shannon, W. Weaver, The mathematical theory of communication. 1949, Urbana, IL: University of Illinois Press.
[65] E. C. Pielou, et al., An introduction to mathematical ecology., An introduction to mathematical ecology.
[66] E. H. Simpson, Measurement of diversity., nature.
[67] A. O. Hirschman, National power and the structure of foreign trade, University of California Press, 1945.
[68] O. C. Herfindahl, Concentration in the us steel industry, Unpublished PhD. Dissertation, Columbia University.
[69] J. P. Gibbs, W. T. Martin, Urbanization, technology, and the division of labor: International patterns, American sociological review (1962) 667–677.
[70] M. Nei, Estimation of average heterozygosity and genetic distance from a small number of individuals, Genetics 89 (3) (1978) 583–590.
[71] S. H. Hurlbert, The nonconcept of species diversity: a critique and alternative parameters, Ecology 52 (4) (1971) 577–586.
[72] A. Sen, M. A. Sen, S. Amartya, J. E. Foster, J. E. Foster, et al., On economic inequality, Oxford University Press, 1997.
[73] W. H. Berger, F. L. Parker, Diversity of planktonic foraminifera in deep-sea sediments, Science 168 (3937) (1970) 1345–1347.
[74] A. R. Solow, S. Polasky, Measuring biological diversity, Environmental and Ecological Statistics 1 (2) (1994) 95–103.
[75] M. L. Weitzman, On diversity, The Quarterly Journal of Economics 107 (2) (1992) 363–405.
[76] P. H. Williams, C. Humphries, R. Vane-Wright, Measuring biodiversity: taxonomic relatedness for conservation priorities, Australian systematic botany 4 (4) (1991) 665–679.
[77] P. Nguyen, P.-P. Saviotti, M. Trommetter, B. Bourgeois, Variety and the evolution of refinery processing, Industrial and corporate change 14 (3) (2005) 469–500.
[78] B. Runnegar, K. Campbell, M. Day, Rates and modes of evolution in the mollusca, Rates of evolution. Allen and Unwin, London (1987) 39–60.
[79] K. Junge, Diversity of ideas about diversity measurement, Scandinavian Journal of Psychology 35 (1) (1994) 16–26.
[80] C. R. Rao, Rao's axiomatization of diversity measures, Wiley StatsRef: Statistics Reference Online.
[81] I. Csiszár, Axiomatic characterizations of information measures, Entropy 10 (3) (2008) 261–273.
[82] G. Wang, M. Jiang, Axiomatic characterization of nonlinear homomorphic means, Journal of Mathematical Analysis and Applications 303 (1) (2005) 350–363.
[83] J. Aczél, C. Alsina, Synthesizing judgements: A functional equations approach, Mathematical Modelling 9 (3-5) (1987) 311–320.
[84] H. Dalton, The measurement of the inequality of incomes, The Economic Journal 30 (119) (1920) 348–361.
[85] S. Hoffmann, Concavity and additivity in diversity measurement: re-discovery of an unknown concept, Univ., FEMM, 2006.
[86] A. N. Kolmogorov, G. Castelnuovo, Sur la notion de la moyenne, G. Bardi, tip. della R. Accad. dei Lincei, 1930.
[87] M. Nagumo, Über eine klasse der mittelwerte, in: Japanese journal of mathematics: transactions and abstracts, Vol. 7, The Mathematical Society of Japan, 1930, pp. 71–79.
[88] H. P. Pursiainen, Consistency in aggregation, quasilinear means and index numbers, Quasilinear Means and Index Numbers (December 4, 2008).
[89] M. Hall, N. Tideman, Measures of concentration, Journal of the american statistical association 62 (317) (1967) 162–168.
[90] A. J. Daly, J. M. Baetens, B. De Baets, Ecological diversity: measuring the unmeasurable, Mathematics 6 (7) (2018) 119.
[91] L. Jost, Entropy and diversity, Oikos 113 (2) (2006) 363–375.
[92] L. Hannah, J. A. Kay, Concentration in modern industry: Theory, measurement and the UK experience, Springer, 1977.
[93] S. Hoffmann, A. Hoffmann, Is there a true diversity?, Ecological Economics 65 (2) (2008) 213–215.
[94] P. Ramaciotti Morales, L. Tabourier, S. Ung, C. Prieur, Role of the website structure in the diversity of browsing behaviors, in: Proceedings of the 30th ACM Conference on Hypertext and Social Media, ACM, 2019, pp. 133–142.
[95] A. Rényi, On Measures of Entropy and Information, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, University of California Press, Berkeley, CA, 1961, pp. 547–561.
[96] T. Van Erven, P. Harremos, Rényi divergence and kullback-leibler divergence, IEEE Transactions on Information Theory 60 (7) (2014)

3797–3820.
[97] S. Kullback, R. A. Leibler, On Information and Sufficiency, The Annals of Mathematical Statistics 22 (1) (1951) 79–86.
[98] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, New York, NY, USA, 2006.
[99] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, AcM, 2008, pp. 1247–1250.
[100] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (11) (1995) 39–41.
[101] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature genetics 25 (1) (2000) 25.
[102] C. Shi, Y. Li, J. Zhang, Y. Sun, S. Y. Philip, A survey of heterogeneous information network analysis, IEEE Transactions on Knowledge and Data Engineering 29 (1) (2016) 17–37.
[103] N. Lao, W. W. Cohen, Fast query execution for retrieval models based on path-constrained random walks, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 881–888.
[104] C. Shi, S. Y. Philip, Heterogeneous information network analysis and applications, Springer, 2017.
[105] M. Kunaver, T. Požrl, Diversity in recommender systems–a survey, Knowledge-Based Systems 123 (2017) 154–162.
[106] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, Y.-C. Zhang, Solving the apparent diversity-accuracy dilemma of recommender systems, Proceedings of the National Academy of Sciences 107 (10) (2010) 4511–4515.
[107] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems, ACM, 2011, pp. 109–116.
[108] G. Salton, M. J. McGill, Introduction to modern information retrieval, mcgraw-hill, 1983.
[109] S. M. McNee, J. Riedl, J. A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in: CHI'06 extended abstracts on Human factors in computing systems, ACM, 2006, pp. 1097–1101.
[110] K. Bradley, B. Smyth, Improving recommendation diversity, in: Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland, Citeseer, 2001, pp. 85–94.
[111] T. Silveira, M. Zhang, X. Lin, Y. Liu, S. Ma, How good your recommender system is? a survey on evaluations in recommendation, International Journal of Machine Learning and Cybernetics 10 (5) (2019) 813–831.
[112] J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez, Recommender systems survey, Knowledge-based systems 46 (2013) 109–132.
[113] R. Burke, Hybrid recommender systems: Survey and experiments, User modeling and user-adapted interaction 12 (4) (2002) 331–370.
[114] A. L'Huillier, S. Castagnos, A. Boyer, Modéliser la diversité au cours du temps pour détecter le contexte dans un service de musique en ligne.
[115] P. Kantor, F. Ricci, L. Rokach, B. Shapira, Recommender systems handbook, Springer, 2010.
[116] J. Tang, S. Wu, J. Sun, H. Su, Cross-domain collaboration recommendation, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 1285–1293.
[117] L. Safoury, A. Salah, Exploiting user demographic attributes for solving cold-start problem in recommender system, Lecture Notes on Software Engineering 1 (3) (2013) 303–307.
[118] G. Groh, C. Ehmig, Recommendations in taste related domains: collaborative filtering vs. social filtering, in: Proceedings of the 2007 international ACM conference on Supporting group work, ACM, 2007, pp. 127–136.
[119] D. Bernardes, M. Diaby, R. Fournier, F. FogelmanSoulié, E. Viennet, A social formalism and survey for recommender systems, Acm Sigkdd Explorations Newsletter 16 (2) (2015) 20–37.
[120] C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th international conference on World Wide Web, ACM, 2005, pp. 22–32.
[121] N. Hurley, M. Zhang, Novelty and diversity in top-n recommendation–analysis and evaluation, ACM Transactions on Internet Technology (TOIT) 10 (4) (2011) 14.
[122] N. Gaumont, M. Panahi, D. Chavalarias, Reconstruction of the socio-semantic dynamics of political activist twitter networksmethod and application to the 2017 french presidential election, PloS one 13 (9) (2018) e0201879.
[123] S. Flaxman, S. Goel, J. M. Rao, Filter bubbles, echo chambers, and online news consumption, Public opinion quarterly 80 (S1) (2016) 298–320.
[124] C. Sha, X. Wu, J. Niu, A framework for recommending relevant and diverse items., in: IJCAI, 2016, pp. 3868–3874.
[125] M. D. Ekstrand, M. Tian, M. R. I. Kazi, H. Mehrpouyan, D. Kluver, Exploring author gender in book rating and recommendation, in: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, 2018, pp. 242–250. doi:10.1145/3240323.3240373.
[126] M. Bastos, D. Mercea, A. Baronchelli, The geographic embedding of online echo chambers: Evidence from the brexit campaign, PloS one 13 (11) (2018) e0206841.
[127] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (6239) (2015) 1130–1132.
[128] C. Roth, Socio-semantic frameworks, Advances in Complex Systems 16 (04n05) (2013) 1350013.
[129] R. T. Paine, Food web complexity and species diversity, The American Naturalist 100 (910) (1966) 65–75.
[130] R. M. May, Ecology: The structure of food webs, Nature 301 (5901) (1983) 566–568.
[131] J. R. Lundgren, Food webs, competition graphs, competition-common enemy graphs, and niche graphs, in: Applications of Combinatorics and Graph Theory to the Biological and Social Sciences, Springer, 1989, pp. 221–243.
[132] R. Poulin, Network analysis shining light on parasite ecology and diversity, Trends in parasitology 26 (10) (2010) 492–498.
[133] D. Urban, T. Keitt, Landscape connectivity: a graph-theoretic perspective, Ecology 82 (5) (2001) 1205–1218.
[134] M. Wirth, G. F. Estabrook, D. J. Rogers, A graph theory model for systematic biology, with an example for the oncidiinae (orchidaceae), Systematic Zoology 15 (1) (1966) 59–69.
[135] C. Kang, Ownership structure of samsung chaebol, Beyond the firm: Business groups in international and historical perspective, ed. T. Shiba, and M. Shimotani (1997) 31–58.

[136] H.-J. Kim, Concentrated ownership and corporate control: Wallenberg sphere and samsung group, J. Korean L. 14 (2014) 39.
[137] H. Compston, The network of global corporate control: implications for public policy, Business and Politics 15 (3) (2013) 357–379.
[138] S. Vitali, J. B. Glattfelder, S. Battiston, The network of global corporate control, PloS one 6 (10) (2011) e25995.
[139] M. Levy, Control in pyramidal structures, Corporate Governance: An International Review 17 (1) (2009) 77–89.
[140] E. Yan, Y. Ding, C. R. Sugimoto, P-rank: An indicator measuring prestige in heterogeneous scholarly networks, Journal of the American Society for Information Science and Technology 62 (3) (2011) 467–477.
[141] M. Zitt, E. Bassecoulard, Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues, Ethics in science and environmental politics 8 (1) (2008) 49–60.
[142] I. Gómez, M. Bordons, M. Fernandez, A. Méndez, Coping with the problem of subject classification diversity, Scientometrics 35 (2) (1996) 223–235.
[143] M. Zitt, Facing diversity of science: A challenge for bibliometric indicators, Measurement: Interdisciplinary Research and Perspectives 3 (1) (2005) 38–49.
[144] G. Heimeriks, M. Hörlesberger, P. Van den Besselaar, Mapping communication and collaboration in heterogeneous research networks, Scientometrics 58 (2) (2003) 391–413.