Abstract

Diversity is a concept relevant to numerous domains of research varying from ecology, to information theory, and to economics, to cite a few. It is a notion that is steadily gaining attention in the information retrieval, network analysis, and artificial neural networks communities. While the use of diversity measures in network-structured data counts a growing number of applications, no clear and comprehensive description is available for the different ways in which diversities can be measured. 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 extends 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 not only provide an effective organization of multiple practices from different domains, but also unearth new observables in systems modeled by heterogeneous information networks. We illustrate the pertinence of our approach by developing 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 social media studies, among other fields.

Keywords: diversity measures, heterogeneous information networks, random walks on graphs, recommender systems, social network analysis.
# Contents

1 Introduction 5

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 9
   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 13
   3.2 Heterogeneous information networks 13
   3.3 Meta paths and constrained random walks 15

4 Network Diversity Measures 16
   4.1 Collective and individual diversities 17
   4.2 Backward diversity 19
   4.3 Relative diversity 20
   4.4 Projected diversity 20
   4.5 The relation between network diversity measures 21
   4.6 Summary of network diversity 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 27
   5.4 Ecology 28
   5.5 Antitrust and competition law 29
   5.6 Scientometrics 30

6 Conclusions 31
1. Introduction

Diversity is a concept of importance in several different domains of research, such as ecology [11], economy [2], public policy [2], information theory [3], social media studies [5, 6], and complex systems [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 including classifications of items, identified as variety (the number of types of entities in the system), 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 Shannon’s entropy in information theory [10], the Gini Index in economy [11], and 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], industrial concentration in economics [14], [15], and online social phenomena such as filter bubbles and echo chambers [5]. The notion of diversity has recently become central as well in the context of digital platforms and online media. The fact that digital platforms increasingly resort to algorithmic recommendations to drive the choices of users has led the scientific community to analyze the impact of recommendations made to users. Although one can argue that this recent development provides users with useful information, the phenomenon also feeds into fears of unpredictable outcomes over the long term, the most debated being the emergence of 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 recommendations on users. On the other hand, many studies have highlighted the need to explore diversity or serendipity (the fortunate discovery of unexpected items) in the way information is exposed to users [21, 22, 23].

Diversity measures can be computed over different types of data in a multitude of contexts. Access to data traces 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 heterogeneous information networks (HINs) [24, 25] have been successfully used 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 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. These sets of paths are called meta paths and can be easily exploited by algorithms. One prominent way of exploiting meta paths is by constraining random walks to them (i.e., constraining random walks to paths contained in a given meta path). This procedure 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], it is gaining widespread use in different communities [40], and in particular in the information retrieval and recommender systems communities [41]. Few studies have hinted at the application of entropy (one prominent diversity measure) to distributions computable from meta path structures in heterogeneous information networks. This application of entropy has been done to provide diverse recommendations [42]. In similarity searches (the search for similar items in information retrieval), 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 from data described with network-structured data. 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 on quantitative diversity measures across different disciplines covering most practical uses. In developing this formal framework, we provide a unified reformulation of several practices existing in scientific literature. In addition, we point to new information that may be extracted by measuring the diversity of previously unconsidered observables in network-structured data. One of the main applications of network diversity measures is the extension of diversity measures, from relatively simple systems of classification and apportionment (e.g., species in ecosystems, units produced by firms) 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 in 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 (especially in applications in recommender systems, information retrieval, and representation learning communities), in particular that of meta paths and observables computable from meta paths;
- the proposal of several network diversity measures, resulting from applying 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 organize diversity measures found in the literature. This framework has the advantage of covering a large part of existing concepts relating to diversity, and 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 diversity measures described within the framework with different observables computed from constrained random walks in order to derive families of interpretable network diversity measures. Finally, in Section 5, we illustrate the relevance of these measures using them in applications in various fields concerned by the concept of diversity.

2. The Concept of Diversity

In general, diversity refers to certain 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 types are from one another. 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), the 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 the way items are classified according to types: item \( i \in I \) is classified as being of type \( t \in T \) if and only if \((i, t) \in \tau\). The use membership relations allows for an item to have more than one type. The diversity measures considered in this article are functions \( D : I \times T \rightarrow \mathbb{R}^+ \) that map any such system to a diversity value \( d \), i.e., \( D : \tau \mapsto d \in \mathbb{R}^+ \).

We do not consider the problem of identification, i.e., what should be considered as an item in a universe of possible elements and what types should be considered in a classification. This identification problem is an important question, however it deals with the meaning of the system’s elements and its semantic content, which is beyond 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 \( p_\tau(t) = \frac{a_\tau(t)}{|I|} \). 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 \( \Delta^k \) to \( \mathbb{R}^+ \), where \( \Delta^k = \cup_{k \geq 0} \Delta^k \) is the union of all standard \( k \)-simplices, that is the set of probability distributions on discrete spaces of size \( k + 1 \):

\[
\Delta^k = \left\{(p_1, \ldots, p_{k+1}) \in \mathbb{R}^{k+1} : \forall i \leq k, \quad p_i \in [0, 1] \text{ and } \sum_{i \leq k+1} p_i = 1\right\}.
\]
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 [45, 46, 47, 1], life sciences [48], economics [2, 12], public policy [49, 50, 51], information theory [3, 4], internet & media studies [5, 6], physics [52, 53], social sciences [54], complexity sciences [55, 56, 7, 8], and opinion dynamics [57]. 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 referred to by the term diversity across the full range of sciences are some combination of three properties, identified as variety, balance, and disparity [9]:

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

The reader is referred to [58] for an extended discussion of these properties.

We illustrate the concept of diversity through classic examples of diversity measures present in work 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** [59, 60] 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, regardless of proportions.

$$R(p) = |\{i \in \{1, 2, \ldots, |T|\} : p_i > 0\}|.$$ Richness only counts types that are effectively used in a classification. If one considers a typology of 20 possible types to examine two bookcases, the first containing books of 3 different types and the second book of 4 different 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 may serve as a basis for other measures, such as, the ratio between richness and the number of classified items [61, Section 9].

**Shannon entropy** [10, 62], here denoted by $H$ and related to the variety and balance properties, is a cross-disciplinary diversity measure, most often used 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 draws books from it at random, Shannon entropy is the average number of binary type-checks (i.e., “yes or no” questions about the book belonging to a given type) one would have to make per book in order to determine its type.

$$H(p) = -\sum_{i=1}^{|T|} p_i \log_2 p_i.$$ Shannon entropy has found renewed use by the information retrieval and artificial intelligence communities. In information retrieval, some recommender systems exploit Shannon entropy to improve performance of algorithmic recommendations [64]. In deep learning methods for artificial intelligence, Shannon entropy is often used for quantifying information gain [65].
The 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 takes 2 books from a bookcase at random, Herfindahl-Hirschman Index is the probability of them belonging to the same type.

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

Also related to the variety and balance properties, this index (also known as the Simpson Index \([65]\)), was first introduced by Hirschman \([67]\) and later by Herfindahl \([68]\) in the study of the concentration of industrial production. Concentration and diversity are opposite and complementary concepts. Higher diversity means lower concentration and vice versa.

A related diversity measure, the Gini-Simpson Index \([11]\) (also called the 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 encounters in ecology \([71]\), it is the probability of the complementary event associated with the Herfindahl-Hirschman Index, i.e., the probability of randomly selecting 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 may be interpreted as a measure of inequality where items are units of wealth distributed into types. One of the formulations of the Gini Coefficient is given by

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

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, regardless of how the remaining 10% of books are classified.

\[
BPI(p) = \max_{n=1,2,...,\|T\|} p_n.
\]

It is easy to see that only the balance property affects this diversity measure. If the books of a first bookcase are classified as 90-10% into two types and those in a second bookcase as 90-5-5% into three types, both bookcases 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 \([74]\). One example of a measure of disparity is proposed in \([75]\):

\[
\text{Disparity}(T) = \sum_{t, t' \in T} d(t, t'),
\]

where \(d\) is a metric on the set \(T\) of types. Disparity is the underlying property in a considerable amount of use cases of use of the notion of diversity. Examples may 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 relies 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, we focus in this article on diversity measures for this latter setting, thus setting aside 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 classification, to non-negative real numbers. Then we further limited the scope to only functions
mapping probability distributions to non-negative real numbers. In this section, we further reduce the scope of diversity measures by prescribing axioms reflecting the 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 these existing axiomatizations, we propose an organization of diversity measures (or true diversities) in the form of an axiom, the resulting measures of the theory correspond to the family of functions known as true diversities. One member of this family, closely related to 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 $D : \Delta^{k-1} \rightarrow \mathbb{R}^+$, 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 types. For instance, a bookcase with 25% comics and 75% novels has the same diversity as a bookcase with 75% comics and 25% novels using a symmetric diversity measure. This means that symmetric diversity measures are blind to the nature of types.

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

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

We also require that diversity measure be expansible, or invariant to non-effective types, that is, invariant to the addition of types with no items. Adding a type with no items does not impact diversity: considering the type “dictionaries” which contains no books does not change the diversity of a bookcase.

Axiom 2 (Expansibility) A diversity measure $D$ is expansible if and only if

$$D(p_1, p_2, ..., p_k) = D(p_1, p_2, ..., p_k, 0).$$

For a diversity measure to be a measure of balance it needs to satisfy the transfer principle, also called the Pigou-Dalton principle [84]: if a bookcase has more novels than comics, replacing some novels with new comics should increase its 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 and only if, for all $i, j$ in $\{1, ..., k\}$, if $p_i > p_j$, then

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

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

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

$$D(\ldots, p_1, p_{i+1}, \ldots) \geq D(\ldots, p_1 + p_{i+1}, \ldots).$$

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 is such that

$$D(1/k, 1/k, \ldots, 1/k) \geq D(p_1, p_2, \ldots, p_k) \geq D(1, 0, \ldots).$$
In order for diversity measures to have a scale for measurement, we need to impose values of 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, its diversity is equal to the number of effective types.

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

$$D(1/k, \ldots, 1/k) = k.$$  

It is easy to see that values of diversity measures 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 may then be used in conjunction with probability computations, as will be shown in Section 4. The measures considered thus far also belong to the more general class of aggregation functions. The most general form of aggregation functions 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 the quantification of information in information theory [3], and are of the form

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

with 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 $\phi$ a strictly monotonic increasing continuous function.

A sub-family of quasilinear means, the so-called self-weighted quasilinear means [88], has additional properties that will be of interest in what 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 $\phi$ 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 simultaneously satisfy the above properties described by the axioms.

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

Theorem 4 provides us with an explicit expression for functions that satisfy axioms 1, 2, 3 & 4. Self-weighted quasilinear means form, however, a subset of the functions defined by these axioms. Indeed, there are diversity measures that satisfy these axioms but are not self-weighted quasilinear means (e.g., 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 \(m\) times, diversity is multiplied by \(m\) \[15\]. Let us suppose, for example, that a bookcase contains 25% comics and 75% novels. Let us also suppose that we add new items from a different bookcase, in which 25% of books are dictionaries and 75% of books are photo albums. The diversity of the new –replicated– bookcase with four types of books is double 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

\[
D \left( \frac{p_1}{m}, \frac{p_2}{m}, \ldots, \frac{p_k}{m} \right) = m \cdot D(1, \frac{p_1}{m}, \ldots, \frac{p_k}{m}).
\]

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

**Definition 2 (True diversity of order \(\alpha\))** The \(\alpha\)-order true diversity, denoted \(D_\alpha\), is the application \(D_\alpha : \Delta^* \rightarrow \mathbb{R}^+\), such that, given \(p = (p_1, \ldots, p_k) \in \Delta^*\) and \(\alpha \in \mathbb{R}^+\),

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

True diversities were first introduced as the Hill Number \[13\] and named true diversity in \[91\]. Variants of true diversities exist in different domains. The Hannah-Kay concentration index of order \(\alpha\) \[22\] is the reciprocal of \(D_\alpha\). In information theory, Rényi Entropy \[4\] of order \(\alpha\), denoted by \(H_\alpha\), is the natural logarithm of \(D_\alpha : H_\alpha(p) = \ln D_\alpha(p)\).

**Theorem 5 (Diversity measures 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 \(\alpha\) if and only if \(D\) satisfies the replication principle of Axiom 5.

True diversities are related to several of the diversities used in different domains and identified in Section 2.3.2. Richness of a distribution \(p\) can be computed as the limit of \(D_\alpha\(1\)(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 0-order true diversity, calling it **Richness diversity**. \(D_1(p)\), 1-order true diversity (or **Shannon diversity**), also called perplexity \[23\], is related to Shannon entropy \(H(p)\) of \(p\) by exponentiation: \(D_1(p) = 2^{H(p)}\) when entropy is computed in base 2. \(D_2(p)\), 2-order true diversity (or **Herfindahl diversity**), is the reciprocal of the Herfindahl-Hirschman Index: \(D_2(p) = 1/\text{HHI}(p)\). The Berger-Parker Index is also 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 \(D_\infty(p) = 1/\text{BPI}(p)\) (here called **Berger diversity**). These relations are summarized in Table 1.

In previous relations, the fact that the Herfindahl-Hirschman Index and the Berger-Parker Index are reciprocal to true diversities underlines the fact that they are intended to measure concentration.

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 \(\alpha\) expresses the way non-uniformity, or balance, is taken into account. If \(\alpha\) is low, inequalities in a distribution will only have a weak impact on diversity values, 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 \(\alpha\) is high, inequalities in a distribution will have a strong impact on
Table 1: Summary of true diversities of order 0, 1, 2, and $\infty$, and their relation to classic diversity measures.

| Order ($\alpha$) | Name                  | True diversity | Expression | Relation to other diversity measures |
|------------------|-----------------------|----------------|------------|--------------------------------------|
| 0                | Richness diversity   | $D_0(p)$      | $\{| i \in \{1, \ldots, k\} : p_i > 0 \}$ | Same as richness [59, 60].          |
| 1                | Shannon diversity    | $D_1(p)$      | $\prod_{i=1}^{k} p_i^{\frac{1}{\alpha}}$ | Exponential of Shannon entropy [10, 62]: $H(p) = \log_2(D_1(p))$, 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]: $HHI(p) = 1/D_2(p)$. |
| $\infty$         | Berger diversity     | $D_{\infty}(p)$ | $\left(\max_{i \in \{1, \ldots, k\}} p_i\right)^{-1}$ | Reciprocal of the Berger-Parker Index [73]: $BPI(p) = 1/D_{\infty}(p)$. |

diversity values, and in the extreme case where $\alpha \to \infty$ (i.e., for Berger diversity), only the highest abundance is taken into account. Red and blue curves in Figure 1 illustrate how parameter $\alpha$ can modulate the relative importance given to variety and balance (cf. Section 2.2): a distribution with 6 types could be less diverse than one with 4 types if it is sufficiently unbalanced for a given value of $\alpha$. True diversities hence allow us to have a continuum of measures which give a different weight to the variety and balance of distributions: $\alpha \to 0$ 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 order $\alpha$, for different distributions.

2.4. Relative true diversities

As with 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 them 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, divergence of the posterior, relative to the prior probability distribution, is a measure of gained information. Relative true diversities generalize this notion using true diversity.
This generalization is analogous to the well-known generalization of the family of Rényi entropies to the family of Rényi divergences \cite{24,53}. Among these generalizations, a well-known special case is the generalization of Shannon entropy to Kullback-Leibler divergence (also known as relative entropy) \cite{96,97}.

Abusing notation, we also denote $D_\alpha$ the $\alpha$-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 $\alpha$ is the 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 $\alpha \in \mathbb{R}^+$,

$$D_\alpha(p \parallel q) = \left( \sum_{i=1}^{k} p_i^\alpha q_i^{1-\alpha} \right)^{\frac{1}{\alpha}} \quad \text{if } \alpha \neq 1.$$

As with true diversities, extreme values are defined as the result of limit processes (cf. Theorems 4, 5, & 6 of \cite{95}):

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

$$D_1(p \parallel q) := \left( \prod_{i=1}^{k} \left( \frac{p_i}{q_i} \right)^{\delta} \right)^{-1} \text{ with } p_i^{\delta} := 1 \text{ if } p_i = 0, \text{ and } D_\infty(p \parallel q) := \max_{i \in \{1, \ldots, k \}} \left( \frac{p_i}{q_i} \right)^{-1}.$$

This definition is analogous to that of true diversities with respect to Rényi entropy: $D_\alpha(p \parallel q) = e^{H_\alpha(p \parallel q)}$. 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 $D_\alpha(u \parallel u) = k/D_\alpha(p)$, and thus $D_\alpha(u \parallel u) \in [1, k]$ (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 $p \in \Delta^{k-1}$, a relative true diversity is only minimal when distributions are equal for all $p, q \in \Delta^{k-1}$

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

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

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 are 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 may be classified according to their genre (e.g., novels, comics) 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 $(i, t_1, t_2) \in \tau_1 \times \tau_2 \iff (i, t_1) \in \tau_1 \land (i, t_2) \in \tau_2$. Let us also define the conditional membership relation $(\tau_2 \mid \tau_1) \subseteq I \times T_2$ such that $(i, t_2) \in (\tau_2 \mid \tau_1) \iff (i, t_2) \in \tau_2 \land (i, t_1) \not\in \tau_1$.

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 $p_{\tau_1} \in \Delta^{|\tau_1|-1}$ and $p_{\tau_2} \in \Delta^{|\tau_2|-1}$. Similarly, we define joint and conditional distributions. We define the joint distribution over $T_1$ and $T_2$ as

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

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

$$p_{(\tau_2 \mid \tau_1)}(t) = \frac{a_{\tau_2 \mid \tau_1}(t)}{|\tau_2 \mid \tau_1|}, \quad \text{for } t_1 \in T_1, \text{ with } p_{(\tau_2 \mid \tau_1)} \in \Delta^{|\tau_2|-1}.$$

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

**Definition 4 (Weak additivity)** A diversity measure $D$ is weakly additive if and only if, for all $\tau_1$ and $\tau_2$ such that $p_{\tau_1 \times \tau_2}(t_1, t_2) = p_{\tau_1}(t_1) p_{\tau_2}(t_2)$, we have $D(p_{\tau_1 \times \tau_2}) = D(p_{\tau_1}) D(p_{\tau_2})$. 

11
Applications from simplices to $\mathbb{R}^+$

Diversities $D$ that satisfy Ax. 1, 2, 3 & 4

Symmetry, Expansibility
Normalisation, Transfer principle
$\Rightarrow$ Merging (Th. 1) & Bounds (Th. 2)

Concave self-weighted quasilinear means

True diversities
Richness
Herfindahl diversity
Berger diversity
Shannon diversity

Replication principle
$\Rightarrow$ Weak additivity (Th. 6)

Strong additivity (Th. 7)

Figure 2: Relations between the different families of diversity measures, and their most important properties.

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 $D_\alpha$ satisfy the principle of weak additivity.

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

A stronger property, called strong additivity principle, and not restricted to independence between $\tau_1$ and $\tau_2$, is verified for the particular case of 1-order true diversity, that is Shannon diversity.

**Definition 5 (Strong additivity)** A diversity measure $D$ is strongly additive if and only if, for all $\tau_1$ and $\tau_2$, we have $D(\pi_{\tau_1 \times \tau_2}) = D(\pi_{\tau_1}) D(\pi_{\tau_2 | \tau_1})$ where $D(\pi_{\tau_2 | \tau_1}) = \prod_{t_1 \in \mathcal{T}_1} D\left(\pi_{\tau_2 | \tau_1}(t_1)\right)$.

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

**Theorem 7 (1-order true diversity is strongly additive [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 [97]): $H(X, Y) = H(X) + H(Y|X)$ for random variables $X$ and $Y$.

Theorem 7 will justify the use of 1-order diversity measures in some results regarding the relations of different network diversity measures in the next sections. Figure 2 summarizes and illustrates the relations between the different families of diversity measures from this section, along with their most important properties.

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 true diversities, which are connected to many of the diversity measures used in different domains of research.
When considering complex systems and network-structured data, different distribution functions can be computed. One example is probability distributions over vertices resulting from a random walk. Different diversity measures may be computed over these distributions. 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 concerned by the concept of diversity.

Developed within graph theory, heterogeneous information networks [24, 25] (equivalent to directed graphs with colored vertices and edges) have recently been used to provide ontologies to represent complex unstructured data in a wide gamut of applications (knowledge graphs are prominent examples of their flexibility [98, 99, 100]). In this work, we will consider an extended model of heterogeneous information network, 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 detail in Section 3, 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 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).

3.1. Preliminary notations

In this article, we consider a multigraph composed of a set of nodes \( V \), linked by a set of directed edges \( E \). We propose the following system of capitalization and typefaces to reference 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 denoted with uppercase letters \( A \), types in \( \mathcal{R} \) are denoted with uppercase letters \( R \);
- **vertex sets** and **edge sets** are notated by uppercase letters \( V \) and \( E \);
- **sets of vertex types** and **edge types** labels are notated by calligraphic letters \( V \) and \( E \);
- random variables with support on sets of vertices by 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 for generality. A multigraph \( G \) is a couple \( (V, E) \) where \( V = \{v_1, \ldots, v_n\} \) is a set of vertices and \( E = \{e_1, \ldots, e_m\} \) is a set of directed edges that is a multisubset of \( V \times V \). Given an edge \( e \in E \), we denote \( v_{\text{src}}(e) \) its source vertex and \( v_{\text{dst}}(e) \) its destination vertex such that \((v_{\text{src}}(e), v_{\text{dst}}(e)) \in V \times V \).

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

- \( e(v_1, v) := \sum_{v_2 \in V} e(v_1, v_2) \) the *out-degree* of vertex \( v_1 \);
- \( e(\cdot, v_2) := \sum_{v_1 \in V} e(v_1, v_2) \) the *in-degree* of vertex \( v_2 \);
- \( e(\cdot, \cdot) := \sum_{(v_1, v_2) \in V \times V} e(v_1, v_2) \) the *total number* of edges.

We now define heterogeneous information networks using multigraphs. Classical heterogeneous information networks can be easily accounted for by constraining the multiplicity of edges.
Definition 6 (Heterogeneous information network) A heterogeneous information network $G = (V, E, A, R, \varphi, \psi)$ is a multigraph $(V, E)$, with a vertex labeling function $\varphi : V \to A$ and an edge labeling function $\psi : E \to R$, such that edges with the same type in $R$ have their source vertices mapped to the same type in $A$ and their destination vertices mapped to the same type in $A$:

$$\forall e, e' \in E, \left( \varphi(e) = \varphi(e') \implies \left( \varphi(v_{\text{src}}(e)) = \varphi(v_{\text{src}}(e')) \land \varphi(v_{\text{dst}}(e)) = \varphi(v_{\text{dst}}(e')) \right) \right).$$

Label functions $\varphi$ and $\psi$, that map 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 $\mathcal{R} = \{R_1, \ldots, R_M\}$, $\varphi$ and $\psi$ induce partitions $V = \{V_1, \ldots, V_N\}$ on $V$ and $E = \{E_1, \ldots, E_M\}$ on $E$. These partitions are such that $\forall v \in V, (\varphi(v) = A_j \iff v \in V_j)$ and $\forall e \in E, (\psi(e) = R_j \iff e \in E_j)$. 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 $\epsilon$ 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 $V_{\text{dst}}(E)$:

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

As before, we also define:

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

Following the example of existing definitions for heterogeneous information networks [24][25][101], we define the network schema. Consistency in the direction of edges belonging to the same edge type allows for the definition of schemas as proper directed graphs. Figure 1 illustrates a heterogeneous information network and its network schema.

Definition 7 (Network schema) The network schema of a heterogeneous information network $G = (V, E, A, R, \varphi, \psi)$ is the directed graph $S = (\mathcal{V}, \mathcal{E})$ 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 $V_{\text{src}}(E)$ is connected to at least one vertex in $V_{\text{dst}}(E)$, i.e., $\forall v_1 \in V_{\text{src}}(E)$ ($\epsilon_E(v_1, -) > 0$), we denote by $p_E : V_{\text{src}}(E) \times V_{\text{dst}}(E) \to [0, 1]$ the transition probability of the random walk following edges in $E$, for all $(v_1, v_2) \in V_{\text{src}}(E) \times V_{\text{dst}}(E)$, 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 $V_{\text{src}}(E)$ to vertex type $V_{\text{dst}}(E)$ in $\mathcal{V}$, we denote the transition from a random vertex $X_{\text{src}} \in V_{\text{src}}(E)$ to a random vertex $X_{\text{dst}} \in V_{\text{dst}}(E)$, following probability distribution $p_E$, as $X_{\text{src}} \xrightarrow{E} X_{\text{dst}}$.

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

In the case where vertex $v_1 \in V_{\text{src}}(E)$ is connected to no vertex in $V_{\text{dst}}(E)$ (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 $e_E^\dagger$ is added such that $v_{\text{src}}(e_E^\dagger)$ is the sink vertex in $V_{\text{src}}(E)$ and such that $v_{\text{dst}}(e_E^\dagger)$ is the sink vertex in $V_{\text{dst}}(E)$. Then, vertices in $V_{\text{src}}(E)$ connected to no vertex in $V_{\text{dst}}(E)$ 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 $V_{\text{src}}(E)$ that are not connected to at least one vertex in $V_{\text{dst}}(E)$.
3.3. Meta paths and constrained random walks

Random walks in heterogeneous information networks can be constrained [34][102] to follow a specific sequence of edge types, called meta path [101][103]. 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 will be the origin of the network diversity measures that this article proposes in the next section.

For the definition of meta paths, we will first consider sequences on the set \( V \) of vertex types. Figure 3 provides an illustration of a heterogeneous information network and a meta path on its network schema.

Definition 10 (Meta path) Given a heterogeneous information network \( \mathcal{G} = (V,E,R,\mathcal{R},\varphi,\psi) \) and a sequence \( r \) of length \( k \) for \( M = |\mathcal{R}| \), a meta path of length \( k \) is the \( k \)-tuple \( \Pi = (E_1, \ldots, E_k) \in \mathcal{E}^k \) of 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 \( \Pi \): i.e., \( \forall 1 \leq i \leq k, V_{src}(E_i) = V_{dst}(E_{i-1}) \).

We denote by \( V_{src}(\Pi) = V_{src}(E_r) \) the source vertex type of path type \( \Pi \), and by \( V_{dst}(\Pi) = V_{dst}(E_r) \) its destination vertex type. Figure 3 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 \( \Pi = (E_1, \ldots, E_k) \) 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 \( \Pi \) 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 \( \Pi \):

\[
X_0 \xrightarrow{E_1} X_1 \xrightarrow{E_2} X_2 \xrightarrow{E_3} \cdots \xrightarrow{E_k} X_k,
\]

where, for all \( i \), \( X_i \in V_{dst}(E_i) \).

This is known as a path-constrained random walk in the information retrieval community [34][102]. It follows from Definitions 8 and 11 that a random walk restricted to a meta path \( \Pi \) of length \( k \) is a Markov chain with transition probabilities defined as

\[
\Pr(X_i = v_i | X_{i-1} = v_{i-1}) = P_{E_i}(v_i | v_{i-1}),
\]

for \( v_{i-1} \in V_{dst}(E_{i-1}) \) and \( v_i \in V_{dst}(E_i) \).

For the next two definitions, we consider a meta path \( \Pi = (E_1, \ldots, E_k) \) of length \( k \) and its associated random walk restricted to \( \Pi \), i.e., the sequence \( (X_0, X_1, \ldots, X_k) \) of random variables. The probability distribution in \( V_{dst}(\Pi) \) of the random walk’s ending vertex plays a 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 $X_k \in V_{dst}(\Pi)$, that is, the destination vertex of the random walk constrained to $\Pi$, given that it started in $v_0 \in V_{src}(\Pi)$ (i.e., $X_0 = v_0$), is denoted by $p_{\Pi}(v_k | v_0)$ for $v_k \in V_{dst}(\Pi)$ and can be recursively computed as follows:

$$p_{\Pi}(v_k | v_0) = \Pr(X_k = v_k | X_0 = v_0) = \sum_{v \in V_{src}(E_{\Pi})} p_{(E_{\Pi} \rightarrow v_1)}(v_k | v_1) p_{E_{\Pi}}(v_1 | v_0).$$

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 conditional probability distribution, the unconditional probability can be computed.

Definition 13 (Unconditional probability distribution for random walks) The unconditional probability distribution of $X_k \in V_{dst}(\Pi)$, that is, the destination vertex of the random walk restrained to $\Pi$, 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 conditional distribution $p_{\Pi|v_0}$ as follows:

$$p_{\Pi}(v_k) = \Pr(X_k = v_k) = \sum_{v \in V_{src}(\Pi)} p_{\Pi|v_0}(v_k) \Pr(X_0 = v_0).$$

In Definition 12, the dependence of $p_{\Pi}$ on $\Pr(X_0 = v_0)$ (the probability distribution for the starting vertex) is explicit.

We now consider the edges resulting from the projection of all edge types in a meta path $\Pi$. This operation, related to the counting of paths in meta paths, is used in the literature on related measures. Counting the number of paths between vertices in a heterogeneous information networks is used in the literature in related measures, such as the construction of similarity metrics for vertex searches [33] or for recommender systems [27].

Definition 14 (Projection of a meta path) Given a meta path $\Pi = (E_1, \ldots, E_\alpha)$, we denote by $E_{\Pi}$ 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 meta path $\Pi$. 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 $\Pi$. It is recursively computed as follows:

$$\epsilon_{E_{\Pi}}(v_0, v_k) = \sum_{v_1 \in V_{src}(E_{\Pi})} \epsilon_{E_1}(v_0, v_1) \epsilon_{E_2, \ldots, E_{\Pi}}(v_1, v_2),$$

with $\epsilon_{(E_1, E_2)} = \epsilon_{E_1}$.

The projection is such that there is an edge in $E_{\Pi}$ for each path in $\Pi$. This allows for the definition of a single step–random walk from $V_{src}(\Pi)$ to $V_{dst}(\Pi)$. Its probability distribution is denoted $p_{E_{\Pi}}$ and computed following Definition 8.

If random walk $X_0 \xrightarrow{E_1} X_1 \xrightarrow{E_2} \cdots \xrightarrow{E_{\Pi}} X_k$ involves choosing a random edge at each vertex type $V_{src}(E_i)$, random walk $X_0 \xrightarrow{E_{\Pi}} X_k$ involves randomly choosing one path among all possible paths in $\Pi$.

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 paths, one can compute several diversities in a network. These diversities will correspond to different concepts related to the structure of vertices and edges in the meta paths: individual, collective, relative, projected, and backward diversity. All of these will be defined in this section. These concepts will in turn have different semantical content 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 definitions and results in this section refer to a heterogeneous information network $G = (V, E, A, R, p, ϕ, ψ)$, and a meta path $Π = (E_r1, . . . , E_rk)$ of length $k$ and going from vertex type $V_{src}(Π)$ to vertex type $V_{dst}(Π)$. Given such a meta path $Π$ and a vertex type $v ∈ V_{src}(Π)$, we define $D_v^Π(Π)$ as the diversity of $v$ along $Π$. The diversity of the vertex type $v$ along $Π$ will depend on the starting probability distribution $Pr(X_0 = v)$ and on transition probabilities $p_{E_{ri}}(v_i | v_{i−1})$ for each $E_{ri} ∈ Π$. Figure 4 provides an example of the measurement of 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 probability distributions $Pr(X_0)$. In Figure 4 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 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 of the random walker’s position during the different steps of the walk).

The choice of $X_0 ∼ Uniform(V_{start})$ has a central role in many applications. By giving each node in $V_{start}$ an equal chance of being the random walk’s starting point, the resulting collective diversity will be that of the collective –equal–contribution of all nodes in $V_{start}$. Similarly, considering subset $V'_{start} ⊂ V_{start}$ and choosing $X_0 ∼ Uniform(V'_{start})$ allows us to define 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 $v_0 ∈ V_{start}$, we define the individual $V_{end}$ diversity of $v_0$ along $Π$ as the true diversity of the probability

\[
D_v\left( X_0 \xrightarrow{Π} X_k \right) = D_v(p_Π). \]

Figure 4: 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 transition probabilities.
Definition 16 (Individual diversity) Given a starting vertex \( v_0 \) of a simple heterogeneous information network, we define the individual diversity of \( v_0 \) as:
\[
D_v(X_0 | X_0 = v_0) = \prod_{v \in V} \frac{1}{P(v | v_0)}
\]

Definition 17 (Mean individual diversity) Given a starting vertex type \( V \), we define the mean \( V \)-individual diversity of \( v_0 \) along \( \Pi \) as the true diversity of the probability distribution of the ending vertex of the constrained random walk. We denote it by:
\[
D_{\Pi}(X_0 \rightarrow X_0 | X_0 = v_0) = D_{\Pi}(p^{(v_0)}).
\]

Figure 5: 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).

Distribution of \( X_k \in V_{end} \) at the end of the constrained random walk, knowing that it started at a given vertex \( v_0 \in V_{start} \) (i.e., \( X_0 = v_0 \)). Figure 5 illustrates the measurement of individual diversity for two different vertices in \( V_{start} \) in the case of 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 \( \Pi \) as the true diversity of the probability distribution of the ending vertex of the constrained random walk. We denote it as \( D_{\Pi}(X_0 \rightarrow X_k | X_0 = v_0) \) and compute it as follows:
\[
D_{\Pi}(X_0 \rightarrow X_k | X_0 = v_0) = D_{\Pi}(p^{(v_0)}).
\]

An aggregation of individual diversities can be computed to represent the mean diversity of all (or many) of the vertices in the starting vertex type \( V_{start} \). Following the definition of conditional entropy in information theory (cf. Section 2.2 in [97]), we define the mean \( V_{end} \)-individual diversity of \( V_{start} \) along \( \Pi \) as the weighted geometric mean of individual diversities.

**Definition 17 (Mean individual diversity)** Given a starting vertex type \( V_{start} \), we define the mean individual \( V_{end} \)-diversity of \( V_{start} \) along \( \Pi \) as the weighted geometric mean of individual diversities. We denote it by:
\[
D_{\Pi}(X_0 \rightarrow X_k | X_0) = \prod_{v \in V_{start}} D_{\Pi}(p^{(v_0)})^{Pr(X_0 = v_0)}
\]

Note that this mean is weighted by \(-\) and so depends 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. Mean individual diversity is a weighted geometric mean in the general case (i.e., for any distribution for \( X_0 \)), and a \(-\)unweighted\(-\) geometric mean when all vertices in \( V_{start} \) have the same probability of being the starting point of the random walk (i.e., when \( X_0 \sim \text{Uniform}(V_{start}) \)). As with collective diversity, the mean individual diversity of a vertex group \( V'_{start} \subset V_{start} \) can be considered choosing \( V'_{start} \sim \text{Uniform}(V'_{start}) \). Figure 5 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 $E^\top$ the set of edges resulting from inverting those of $E$:

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

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

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

Using random walks along a meta path $\Pi$, we can also compute the probability distribution of the random walk’s starting vertex $X_0 \in V_{\text{start}}$ when its ending vertex $v_k \in V_{\text{end}}$ is known. True diversity of this distribution, called backward $V_{\text{start}}$ diversity of $v_k \in V_{\text{end}}$ along $\Pi$, provides a value for the diversity of starting points that can reach $v_k$ following $\Pi$.

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

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

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

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

Definition 21 (Relative individual diversity) We define the relative individual $D_{v_0}$ diversity of $v_0 \in V_{\text{start}}$ with respect to $V_{\text{start}}$ along $\Pi$ as the relative true diversity between the distribution resulting from a random walk starting at $v_0 \in V_{\text{start}}$ (giving its individual diversity), and the distribution resulting from the unconditional random walk starting at random in $V_{\text{start}}$ (giving the collective diversity). We denote it by $D_{\alpha} \left(X_0 \xrightarrow{\Pi} X_k | X_0 = v_0 \| X_0 \xrightarrow{\Pi} X_k \right)$ and compute it as follows:

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

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

$$D_{\alpha} \left(X_0 \xrightarrow{\Pi_1} X_k | X_0 = v_0 \| X_0 \xrightarrow{\Pi_2} X_k \right) = D_{\alpha} \left(p_{\Pi_1} | p_{\Pi_2} \right).$$

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

4.4. Projected diversity

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

Definition 22 (Projected diversity) Let $E_{\Pi}$ be the set of projected edges of meta path $\Pi$. We define the projected $V_{\text{end}}$ diversity of $v_0 \in V_{\text{start}}$ along $\Pi$ as the true diversity of the distribution of the ending vertices in $V_{\text{end}}$ of a random walk starting at $v_0 \in V_{\text{start}}$ and following the edges in $E_{\Pi}$. We denote it by $D_{\alpha} \left(X_0 \xrightarrow{E_{\Pi}} X_k | X_0 = v_0 \right)$ and compute it as follows:

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

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 projected edges in $E_{\Pi}$. Figure 7 illustrates the comparison between individual and projected diversities for two cases using 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.

4.5. The relation between network diversity measures

The network diversity measures presented here are not independent. In this section we show a relation involving collective, backward, and mean individual diversities. In order to do so, we first need to consider parts of meta paths.

Definition 23 (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

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

Theorem 8 (Bound for collective Shannon diversity) The following inequality holds for Shannon diversity, that is 1-order true diversity,

$$D_{\alpha} \left(X_0 \xrightarrow{E_{r_i}} X_k \right) \leq D_{\alpha} \left(X_0 \xrightarrow{\Pi_{(i,j)}} X_k \right) \leq D_{\alpha} \left(X_0 \xrightarrow{\Pi_{(i,j)}} X_k \right)$$

with equality if and only if $D_{\alpha} \left(X_0 \xrightarrow{E_{r_i}} X_k \right) = 1$. 

20
In other words, collective diversity along a meta path is bounded by two factors: (1) collective diversity at any step of the meta path, multiplied by (2) mean individual diversity along the remaining part of the meta path. This bound is achieved if and only if all individual backward diversities along this remaining part are minimal.

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

**Lemma 1 (The 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:

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

where $D_1 \left( \Pr(X_0) \right)$, the initial diversity, 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 in $V_0$ and the ending vertex in $V_1$. Despite $X_0$ and $X_1$ being dependent, by the principle of strong additivity of 1-order true diversity (cf. Theorem 7), we have

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

Figure 7: Individual Shannon diversities of a meta path on two heterogeneous information networks, compared with the resulting projected diversities.
Also by the principle of strong additivity of 1-order true diversity we have

$$D_1(\Pr(X_0, X_1)) \equiv D_1(\Pr(X_1)) \prod_{v_i \in V_i} D_1(\Pr(X_0 | X_1 = v_i))^{\Pr(X_i = v_i)}$$


Def. 20

$$D_1 \left(X_0 \xrightarrow{E} X_1 \right) = D_1 \left(X_0 \xrightarrow{E} X_1 | X_0 \right)$$

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

$$D_1 \left(X_0 \xrightarrow{E} X_1 \right) \leq D_1(\Pr(X_0)) \prod_{v_i \in V_i} D_1(\Pr(X_0 | X_1 = v_i))^{\Pr(X_i = v_i)}$$

with equality when mean backward diversity is minimal, $D_1 \left(X_0 \xrightarrow{E} X_1 \right) = 1$. This can only happen when each ending vertex in $V_i$ 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:

$$\Pi_{(1,0)} = (E_{r_1}, \ldots, E_{r_v})\text{, for random walk } X_0 \xrightarrow{\Pi_{(1,0)}} X_v$$

$$\Pi_{(v+1,k)} = (E_{r_{v+1}}, \ldots, E_{r_k})\text{, for random walk } X_v \xrightarrow{\Pi_{(v+1,k)}} X_k$$

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

A first application of the strong additivity principle yields

$$D_1(\Pr(X_v, X_k)) = D_1 \left(X_0 \xrightarrow{\Pi_{(1,0)}} X_v \right) \sum_{\Pi_{(v+1,k)}} D_1 \left(X_v \xrightarrow{\Pi_{(v+1,k)}} X_k \right)$$

where $D_1 \left(X_v \xrightarrow{\Pi_{(v+1,k)}} X_k \right)$ is the mean individual diversity along meta path $\Pi_{(v+1,k)}$ using probabilities resulting from random walk $X_0 \xrightarrow{\Pi_{(1,0)}} X_k$ for the weighted geometric mean.

A second application of the strong additivity principle yields

$$D_1(\Pr(X_v, X_k)) = D_1 \left(X_v \xrightarrow{\Pi_{(v+1,k)}} X_k \right) \sum_{\Pi_{(1,0)}} D_1 \left(X_v \xrightarrow{\Pi_{(1,0)}} X_k \right)$$

Since starting probabilities $\Pr(X_v)$ in collective diversity $D_1 \left(X_v \xrightarrow{\Pi_{(v+1,k)}} X_k \right)$ are those resulting from random walk $X_0 \xrightarrow{\Pi_{(1,0)}} X_v$, we also have $D_1 \left(X_v \xrightarrow{\Pi_{(v+1,k)}} X_k \right) = D_1 \left(X_0 \xrightarrow{\Pi} X_k \right)$.

This gives the desired result

$$D_1 \left(X_0 \xrightarrow{\Pi} X_k \right) = D_1 \left(X_0 \xrightarrow{\Pi_{(1,0)}} X_v \right) \sum_{\Pi_{(v+1,k)}} D_1 \left(X_v \xrightarrow{\Pi_{(v+1,k)}} X_k \right)$$

from which it follows that

$$D_1 \left(X_0 \xrightarrow{\Pi} X_k \right) \leq D_1 \left(X_0 \xrightarrow{\Pi_{(1,0)}} X_v \right) \sum_{\Pi_{(v+1,k)}} D_1 \left(X_v \xrightarrow{\Pi_{(v+1,k)}} X_k \right)$$

if mean backward diversity is not equal to 1. ■
4.6. Summary of network diversity measures

In this section, we have used the definitions developed within the proposed formalism for heterogeneous information networks, in particular that of meta path constrained random walk, to propose different network diversity measures. These include collective, individual, backward, relative, and projected diversities along a meta path. For each one, we have proposed a notation and we have defined a 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 present different domains of application for which modeling using heterogeneous information networks is useful. We show that some quantitative measures traditionally computed in diverse domains are closely related to the network diversity measures we defined, and that their use allows for the consideration of other useful quantitative observables in modeled systems.

5. Applications

The network diversity measures we proposed find numerous applications in many domains where diversity provides relevant information. Information retrieval, and in particular algorithmic recommendation, is one of the areas with the most direct applications that best illustrates 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. Recommender systems, closely related to information retrieval in computer sciences, intersects with many research topics in artificial intelligence, machine learning, and data mining, and will help us provide illustrative examples for these domains. The first example introduces notations used in this section, and the approach chosen to illustrate the application of these measures. This approach consists of 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 in each domain, we list several concepts of interest for research questions that are traditionally relevant in the literature, together with the explicit expression of the corresponding network diversity measures. These concepts will be referred to previous work where they find pertinence. We highlight how these proposed measures can address existing research questions and current practices in different research areas, and how they allow for positing new ones.

After having introduced a simple first example, we provide a more detailed application case in a setting from recommender systems in Section 5.2 explaining the relation between network diversity measures and several existing practices and concepts while also highlighting possible new uses. We then illustrate the use of network diversity measures for the analysis of social networks and media in Section 5.3. Finally, we provide other examples of applications in ecology in Section 5.4, antitrust regulation in Section 5.5 and scientometrics in Section 5.6.

5.1. A simple example

Let us consider an example heterogeneous information network with three vertex types: users, items, and types of items. Similar to the notation established in Section 5.1 for the sake of readability, let us denote these vertex types respectively by $V_{\text{users}}, V_{\text{items}},$ and $V_{\text{types}}$. An example of entities represented by items are films, and an example of types are then film genres (e.g. comedy, thriller).

| Diversity       | Notation | Expression |
|-----------------|----------|------------|
| Collective      | $D_{\alpha}(X_0 \xrightarrow{p_0} X_k)$ | $D_{\alpha}(p_{0i0})$ |
| Individual      | $D_{\alpha}(X_0 \xrightarrow{p} X_k | X_0 = v_0)$ | $D_{\alpha}(p_{0ii0})$ |
| Mean individual | $\prod_{v_0 \in V_0} D_{\alpha}(p_{0ii0})$ | $\prod_{v_0 \in V_0} D_{\alpha}(p_{0ii0})$ |
| Relative individual | $D_{\alpha}(X_0 \xrightarrow{p} X_k | X_0 = v_0 ; X_k \xrightarrow{p} X_k)$ | $D_{\alpha}(p_{0ii0} || p_{1i1})$ |
| Backward individual | $D_{\alpha}(X_0 | X_0 \xrightarrow{p} X_k)$ | $D_{\alpha}(p_{0ii0})$ |
| Projected individual | $D_{\alpha}(X_0 \xrightarrow{p} X_k | X_0 = v_0)$ | $D_{\alpha}(p_{0ii0})$ |
Let us now consider three edge types, indicating items chosen by users, items recommended to users, and classification of items into types. We respectively denote these edge types as $E_{\text{chosen}}$, $E_{\text{recommended}}$, and $E_{\text{types}}$. Figure 8 illustrates the network schema of the described heterogeneous information network.

![Figure 8: Network schema of a simple heterogeneous information network, where users in $V_{\text{users}}$ have chosen and have been recommended items in $V_{\text{items}}$, which are classified into types in $V_{\text{types}}$.]

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, $X_{\text{users}}$ is a random vertex in $V_{\text{users}}$, i.e., a random user. This allows for the consideration of 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, 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),

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

that quantifies the type diversity of items that are recommended to the users. We might also take interest in the mean individual type diversity of items recommended to users (cf. Definition 17),

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

that quantifies the mean of the type diversity of items recommended to each user. Network diversity measures allow for the evaluation, for example, of the collective type diversity of items that are recommended to users with respect to items that are chosen by users using relative diversity (cf. Definition 21):

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

Such a measure would reveal how diverse recommendations are (according to item’s types) while taking the general landscape of users’ consumption as a baseline to measure this diversity. In other words, such a measure would reveal how recommendations may increase or decrease the diversity of what is consumed.

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

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

which would otherwise be referred to as the individual type diversity of items chosen by users that chose items recommended to user $u \in V_{\text{users}}$. Some random variables are marked with an apostrophe, e.g., $X'_{\text{users}}$, 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 use this approach: to identify, referenciate, and provide computable expressions for concepts from different domains of research interested in both diversity measures and network representations.

24
5.2. Recommender Systems

Diversity and diversification of algorithmic recommendations has become one of the leading topics of the recommender systems research community [104] [105]. 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 all that is offered. Recommender systems, developed as early as in the 1980s [106], help solve this problem by filtering all 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 of diversity and its introduction in recommendations [107] [108].

In recommender systems, diversity can help improve users’ appreciation of the quality of recommendations [109] [110]. It also has other applications, such as detecting changes in consumption behavior for context-aware recommenders [111]. 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 [112], or [109] [110] for a discussion of terminology and definitions). These indicators are often computed using past collective choices of items made by users [105], or classifications of items into types [20]. To this date, no general framework exists to account for all proposed diversity indices in recommender systems, nor alternatives for exploiting richer meta-data structures such 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, extend the measurement of diversity to more complex data structures that can include meta-data on users and items, and give formal explicit expressions to computable quantities related to new and existing research questions in this field.

For illustrative purposes, 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 9 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:

- 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 9: 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 $V_{T_2}$.](image)

In order to consider random walks constrained to meta paths, and following the example in 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 Figure 9, we also consider different edge types:
• An edge type between users $E_{U}$ (e.g., users following or friending each others on a social network);
• Edge types from groups to users: $E_{G_{i}}$ (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: $E_{r_{11}}$ and $E_{r_{12}}$ for $V_{I_{1}}$, and $E_{r_{22}}$ for $V_{I_{2}}$;
• Edge types representing when users have liked ($E_{like}$), seen ($E_{seen}$), or rated ($E_{rate}$) an item, or representing when users have been recommended items ($E_{rec,1}$ and $E_{rec,2}$).

All 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 (explicit feedback framework, this concept can be captured by the individual diversity. Let us imagine that $V$ can be computed, for example, with respect to a classification of items [118] (e.g., genres for films) . In the proposed framework an aggregation of this quantity for all users. Diversity between items can be also exploited for recommendations [116, 117], and certainly in diversity computations. As stated before, Figure 9 represents the network schema (cf. Definition 7) of our example.

Most diversity computations in recommender systems are concerned with providing a measure of the diversity of 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 items [118] (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 $V_{I_{1}}$ are films, and that $V_{I_{2}}$ are film genres (e.g., comedy, thriller, etc.). The individual genre ($V_{I_{2}}$) diversity of films ($V_{I_{1}}$) recommended to a user $u \in V_{U}$ is

$$D_{\alpha}(X_{U} \xrightarrow{E_{rec,2}} X_{I_{2}} \xrightarrow{E_{like}} X_{I_{1}} \mid X_{U} = u).$$

Similarly, the mean genre ($V_{I_{2}}$) diversity of films ($V_{I_{1}}$) recommended to all users $V_{U}$ is the mean individual diversity

$$D_{\alpha}(X_{U} \xrightarrow{E_{rec,2}} X_{I_{2}} \xrightarrow{E_{like}} X_{I_{1}} \mid X_{U}),$$

which can be computed as a geometric mean by choosing $X_{U} \sim \text{Uniform}(V_{U})$ for the starting point 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 seen or liked it (sometimes called novelty [119]). In the proposed framework, an aggregation of this quantity for items proposed to all users corresponds to the following network diversity measure:

$$D_{\alpha}(X_{U} \xrightarrow{E_{rec,1}} X_{I_{1}} \xrightarrow{E_{like}} X_{I_{2}} \mid X_{U}).$$

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, $D_{\alpha}(X_{U} \xrightarrow{E_{rec,1}} X_{I_{1}} \xrightarrow{E_{like}} X_{I_{2}})$ versus $D_{\alpha}(X_{U} \xrightarrow{E_{rec,1}} X_{I_{1}} \xrightarrow{E_{like}} X_{I_{2}}).$

Distinguishing between 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 6). Other concepts expressible as network diversity measures are related to the computations underlying recommender systems such as the so-called User-Based Collaborative Filtering:

$$D_{\alpha}(X_{U} \xrightarrow{E_{rec,1}} V_{I_{1}} \xrightarrow{E_{like}} X_{U} \xrightarrow{E_{like}} V_{I_{1}} \xrightarrow{E_{rec,2}} V_{I_{1}}).$$
which corresponds to the collective type diversity of items chosen by users that chose items recommended to users.

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 systems, along with the respective quantities that can be identified, expressed, and computed as network diversity measures.

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 measures |
|------------------------------------------------------|----------------------------------------|
| Mean individual diversity of recommendation of items $V_{11}$ according to types $V_{12}$ relative to the corresponding collective diversity | $D_0 \left( X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \right)$ |
| Collective diversity of recommended items $V_{11}$ according to types $V_{12}$ relative to the distribution of types of liked times | $D_0 \left( X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \right)$ |
| Collective diversity of recommendations of items in $V_{11}$ (e.g. films) according to types $V_{12}$ (e.g. genres) relative to the one of $V_{12}$ (e.g. series) | $D_0 \left( X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \right)$ |
| Diversity of items $V_{11}$ recommended to friends of $u \in V_c$ according to types $V_{12}$ | $D_0 \left( X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \right)$ |
| Diversity of items $V_{11}$ liked by group $g \in V_{c_1}$ according to types $V_{12}$ | $D_0 \left( X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \right)$ |
| Diversity of users $V_c$ that liked items $V_{11}$ of type $v \in V_{c_2}$ | $D_0 \left( X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \right)$ |
| Diversity of types $V_{12}$ chosen by user $u \in V_c$ through their choices of items in $V_{11}$ | $D_0 \left( X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \xrightarrow{\mu_{rec}} X_1 \xrightarrow{\mu_{rec}} X_1 \parallel X_0 \right)$ |

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 on social networks has become crucial. Heterogeneous information networks provide 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 or following other users, using tags in publication). More complex and abstract data is often analyzed in these studies, such as the political affiliations of users and media outlets [120][121]. 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][122][123][23]. The settings of different social media studies vary. Concrete examples are the study of the Leave and Remain Brexit campaigns on Twitter [124] and the exchange of information between US Democrats and Republicans on Facebook [125].

In this section, we illustrate the use of network diversity measures for the study of information exchange on social networks. In order to do so, we consider, as before, a heterogeneous information network for illustration, now in the setting of activity traces of social networks and media. Figure 10 shows a network schema chosen to illustrate the applicability of the proposed network diversity measures to this context. The heterogeneous information network for this example considers: users that post or share posts (or tweets, or blog entries, or comments in forums), users that can follow (or friend) 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), and even link to articles through a URL address. In many contexts, articles can be associated with media outlets, which can in turn be identified with groups or affiliations (e.g., political parties).

The considered heterogeneous information network can accommodate different aspects considered in social media studies. For example, Gaumont et al. [120] consider relations of political affiliation of users and interactions between them, and analyze 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 [126], explicitly consider networks of information items (e.g., blog posts) and the concepts that they use.

As with the example of a generic recommender systems, we present in a schematic fashion (in Table 3) different diversity-related concepts of interest for research questions in the field of social networks and media studies, along with quantities that can be computed as network diversity measures.
Figure 10: 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.

Table 4: Schematic representation of diversity-related concepts 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 measures |
|--------------------------------------------------------|-----------------------------------------|
| Collective diversity of affiliations of users           | $D_\alpha (X_{users} \mapsto X_{affiliations})$ |
| Collective affiliation diversity of users through the contents they share | $D_\alpha (X_{users} \mapsto X_{posts} \mapsto X_{articles} \mapsto X_{medias} \mapsto X_{affiliations})$ |
| Individual topic diversity of user $u \in V_{users}$ through posting | $D_\alpha (X_{users} \mapsto X_{posts} \mapsto X_{topics} | X_{users} = u)$ |
| Individual topic diversity of user $u \in V_{users}$ through posting of followers | $D_\alpha (X_{users} \mapsto X_{posts} \mapsto X_{topics} | X_{users} = u)$ |
| Individual affiliation diversity of users mentioned by user $u \in V_{users}$ | $D_\alpha (X_{users} \mapsto X_{posts} \mapsto X_{topics} | X_{users} = u)$ |
| Diversity of affiliation groups that treat topic $t \in V_{topics}$, in articles | $D_\alpha (X_{affiliations} \mapsto X_{affiliations} \mapsto X_{medias} \mapsto X_{articles} \mapsto V_{topics} = t)$ |
| Affiliation diversity of users according to who they mention in their posts relative to their own identified political affiliation | $D_\alpha (X_{users} \mapsto X_{posts} \mapsto X_{topics} | X_{users} = u)$ |

5.4. Ecology

Diversity is useful in ecology, as identified and commented in Section 2.2. Many advances in diversity measures come from this community (e.g., [13]). One prominent concept in this domain is the diversity of species in a 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 a habitat, diversity is then computed and reported.

Interactions among organisms are also of interest in ecology. These can be treated using graph representations and models. One of such interactions, also related to diversity, is represented by so-called food webs [127]: network models that describe species that feed on other species. In the past, there have been efforts to use graph formalisms to treat food webs [128, 129]. Similarly, other relations between species have been described using graphs, such as parasitisation [130]. Another subject of interest in ecology is the description of habitats and their interconnectedness; there have been several approaches using graph theory to describe these connections [131, 132].
All these elements of interest in ecology can be treated using heterogeneous information networks. Hence, using network diversity measures, different concepts of interest related to diversity can be computed. Let us consider for example a heterogeneous information network with vertex types for habitats ($V_{\text{habitats}}$), for individuals ($V_{\text{individuals}}$), for species ($V_{\text{species}}$), for genera ($V_{\text{genera}}$), for families ($V_{\text{families}}$), and so on as needed.

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

This setting can accommodate the common practices of measurement of –bio– diversity of a habitat

\[
D_\alpha \left( X_{\text{habitats}} \xrightarrow{E_{\text{habitu}}} X_{\text{individuals}} \xrightarrow{E_{\text{habitu}}} X_{\text{species}} \big| X_{\text{habitats}} = h \right),
\]

with $\alpha = 0$ giving the richness biodiversity, and $\alpha = \infty$ the Berger-Parker biodiversity of the habitat. Figure 11 illustrates the network schema of the described heterogeneous information network, along with a table of diversity-related concepts and their expressions as network diversity measures.

![Network schema of an example from ecology, and table with examples of diversity-related concepts and their expression as network diversity measures.](image)

**Table 11:** Network schema of an example from ecology, and table with examples of diversity-related concepts and their expression as network diversity measures.

**5.5. Antitrust and competition law**

Many developments and applications of concentration measures are found in the economics community, antitrust regulation, and competition law. As shown in Section 2.2, concentration is a concept for which indices are the
reciprocal of those used for the concept of diversity. 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 tons of steel produced—in a given period of time in a given country—by the firms that produced them. From this apportionment or distribution, concentration of the steel industry can be quantitatively measured 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 concentration of an industry allows for important comparisons to be made by industry regulators, such as, for example, the degree of concentration of an industry should a given merger or 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 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 can name board members of other firms in the same industry producing complex relations of control in a network-like fashion.

Specialized economics literature accounts for many case studies that challenge the application of the aforementioned procedure to regulation [133][134], and that address the complex structure of co-ownership networks and their importance in regulation [135][136]. This makes graph-theoretical approaches good candidates for making advances in the measurement of concentration in industries [137]. In particular, the proposed network diversity measures provide tools that allow the measurement of many concepts of interest in antitrust regulation and competition law when dealing with network structures.

To illustrate this, let us consider 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 of 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 units to the firm that produced them, $E_{own}$, linking firms to the persons that own them, $E_{cross-own}$, linking firms with each other according to cross-ownership, and similarly, $E_{control}$ and $E_{cross-control}$ linking firms with each other according to control and cross-control (for example, having the right to choose a member of the board of a given firm). For edge types $E_{own}$, $E_{cross-own}$, $E_{control}$, and $E_{cross-control}$, the multiplicity of edges can account for the units by which property or control is represented, such as, for example, shares or members of the boards of the firms. For example, if ownership of a firm is represented by 10 shares, it will have 10 edges, that can belong to edge types $E_{own}$ or $E_{cross-own}$.

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

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

which becomes the Herfindahl-Hirschman Diversity (reciprocal of the Herfindahl-Hirschman Index) by choosing $\alpha = 2$. Figure[12] illustrates the heterogeneous information network described in this example, along with a table of concepts related to diversity and expressible using the proposed network diversity measures.

5.6. Scientometrics

Scientometrics, within the field of bibliometrics, studies the measurement and analysis of scientific literature. Overlapping 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 measurements [138].

The study of networks modeling and representing scientific production is of interest for other reasons too. Diversity of topics explored by scientific communities is a concept of interest, for example, in public policy [139], and in general for the understanding and description of the structure of scientific communities [140][141]. Another practical application of the measurement of diversity in citation networks is the maintenance of classification systems [142].
Examples of concepts expressible in research questions

| Concept                                               | Corresponding network diversity measures |
|-------------------------------------------------------|------------------------------------------|
| Industry diversity with cross-ownership relations     | $D_\alpha(X_{\text{units}} \xrightarrow{E_{\text{produced}}} X_{\text{firms}} \xrightarrow{E_{\text{cross-own}}} Y_{\text{firms}} \xrightarrow{E_{\text{control}}} X_{\text{persons}})$ |
| Industry diversity according to persons with cross-ownership relations | $D_\alpha(X_{\text{units}} \xrightarrow{E_{\text{produced}}} X_{\text{firms}} \xrightarrow{E_{\text{cross-own}}} Y_{\text{firms}} \xrightarrow{E_{\text{control}}} X_{\text{persons}} \xrightarrow{E_{\text{own}}} X_{\text{persons}})$ |
| Diversity of cross-ownership of a firm $f \in V_{\text{firms}}$ | $D_\alpha(X_{\text{firms}} \xrightarrow{E_{\text{cross-own}}} Y_{\text{firms}} \mid X_{\text{firms}} = f)$ |
| Diversity of ownership of firms relative to their diversity of control | $D_\alpha(X_{\text{firms}} \xrightarrow{E_{\text{own}}} X_{\text{persons}} \parallel Y_{\text{firms}} \xrightarrow{E_{\text{control}}} X_{\text{persons}})$ |

Figure 12: 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 by each others.

In this section, we illustrate the way 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 consisting of the following vertex types: authors $V_{\text{authors}}$, laboratories $V_{\text{lab.}}$ (or affiliation institutions), journals $V_{\text{journals}}$, scientific articles $V_{\text{papers}}$, keywords used by these articles $V_{\text{keywords}}$, and domains of research $V_{\text{domains}}$ (e.g., ecology, economics). We also consider edge types for representing relations between these entities (see Figure 13): affiliation $E_{\text{aff}}$ of authors to institutions, edition (or peer-review) $E_{\text{edit}}$ of journals by authors, writing $E_{\text{write}}$ of articles by authors, use of keywords $E_{\text{use}}$ in articles, association of keywords $E_{\text{belong}}$ with domains, publication $E_{\text{publish}}$ of articles by journals, and declared treatment $E_{\text{treat}}$ of research domains by journals. Figure 13 illustrates the corresponding heterogeneous information network, along with a table 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 in 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 $\alpha$: 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 phenomena or objects 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 in 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 applications in several domains.

The main contributions of this article are:

- The proposition of an axiomatic theory of diversity measures that allows us to present most of their uses across several domains with a single-parameter family of functions.
Examples of concepts expressible in research questions

| Corresponding network diversity measures |
|------------------------------------------|
| Diversity of keywords used by author \( a \in V_{authors} \) |
| \( D \alpha(X_{authors} \xrightarrow{\text{write}} X_{papers} \xrightarrow{\text{use}} X_{keywords} | X_{authors} = a) \) |
| Diversity of domains addressed in publications by author \( a \in V_{authors} \) relative to domains he or she addresses in editing (or peer-reviewing) |
| \( D \alpha(X_{authors} \xrightarrow{\text{write}} X_{papers} \xrightarrow{\text{use}} X_{keywords} \xrightarrow{\text{belong}} X_{domains} | X_{authors} = a \| X_{\text{edit}} \) |
| Diversity of domains addressed by citations by authors of laboratory \( l \in V_{lab} \) |
| \( D \beta(X_{\text{lab}} \xrightarrow{\text{aff}} X_{authors} \xrightarrow{\text{write}} X_{papers} \xrightarrow{\text{cite}} X_{\prime papers} \xrightarrow{\text{use}} X_{keywords} \xrightarrow{\text{belong}} X_{domains} | X_{\text{lab}} = l \) |

Figure 13: Network schema of a heterogeneous information network in an example for scientometrics, and table of examples of concepts related to diversity and expressible using the proposed network diversity measures.

- The formalization of concepts and tools to describe and process heterogeneous information networks, that have been gaining attention in 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 probability distributions that are computable with 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 pre-existing quantitative measurements that are widespread in different fields, and the development of new applications through examples in recommender systems, social media studies, ecology, competition law, and scientometrics.

We hope that this framework for the application of diversity measures to network structures will enrich research on diversity in the domains identified in Section 5 and beyond. Future developments in this line of research might consider the identification of algebraic structures for network diversity measures, and their application to case studies.

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).

The authors would like to sincerely thank Hông-Lan Botterman and Matthieu Latapy for their comments, as well as Claire Schaffer for her thorough proof-reading.
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. Kulisbresha, M. B. Zafar, L. E. Nobou, K. P. Gummmadi, S. Ghosh, Characterizing information diets of social media users, in: ICWSM, 2015, pp. 218–227.
[7] R. K. Uremse, 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 arpsso, 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 notion and its consequences, Ecology 54 (2) (1973) 427–432.
[14] E. 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. Tschante, Discrimination in online personalization: A multidisciplinary inquiry, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, ACM, 2018.
[19] R. Courtland, Bias defaults: The researchers striving to make algorithms fair, Nature 558 (7710) (2018) 357–360.
[20] C.-N. Ziegler, S. M. McNeel, J. A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th international conference on Web search and data mining, 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. Schell, P. Keeses, 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.-Y. Xu, 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. Piikula, R. Schenkkel, 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] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, A. Arenas, Ranking in inter-connected multilayer networks reveals versatile nodes, Nature communications 6 (2015) 6868.
[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.
[121] S. Flaxman, S. Goel, J. M. Rao, Filter bubbles, echo chambers, and online news consumption, Public opinion quarterly 80 (S1) (2016) 298–320.

[122] C. Sha, X. Wu, J. Niu, A framework for recommending relevant and diverse items., in: IJCAI, 2016, pp. 3868–3874.

[123] M. D. Ekstrand, M. Tian, M. R. I. Kazi, H. Mehrpouyan, D. Klüver, 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

[124] M. Bastos, D. Mercea, A. Baronchelli, The geographic embedding of online echo chambers: Evidence from the brexit campaign, PloS one 13 (11) (2018) e0206841.

[125] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (6239) (2015) 1130–1132.

[126] C. Roth, Socio-semantic frameworks, Advances in Complex Systems 16 (04n05) (2013) 1350013.

[127] R. T. Paine, Food web complexity and species diversity, The American Naturalist 100 (910) (1966) 65–75.

[128] R. M. May, Ecology: The structure of food webs, Nature 301 (5901) (1983) 566–568.

[129] 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.

[130] R. Poulin, Network analysis shining light on parasite ecology and diversity, Trends in parasitology 26 (10) (2010) 492–498.

[131] D. Urban, T. Keitt, Landscape connectivity: a graph-theoretic perspective, Ecology 82 (5) (2001) 1205–1218.

[132] M. Wirth, G. F. Estabrook, D. J. Rogers, A graph theory model for systematic biology, with an example for the oncidinae (orchidaceae), Systematic Zoology 15 (1) (1966) 59–69.

[133] C. Kang, Ownership structure of samsung chaebol, Beyond the firm: Business groups in international and historical perspective, ed. T. Shiba, and M. Shimotani (1997) 51–58.

[134] H.-J. Kim, Concentrated ownership and corporate control: Wallenberg sphere and samsung group, J. Korean L. 14 (2014) 39.

[135] H. Compston, The network of global corporate control: implications for public policy, Business and Politics 15 (3) (2013) 357–379.

[136] S. Vitali, J. B. Glattfelder, S. Battiston, The network of global corporate control, PloS one 6 (10) (2011) e25995.

[137] M. Levy, Control in pyramidal structures, Corporate Governance: An International Review 17 (1) (2009) 77–89.

[138] 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.

[139] 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.

[140] M. Zitt, Facing diversity of science: A challenge for bibliometric indicators, Measurement: Interdisciplinary Research and Perspectives 3 (1) (2005) 38–49.

[141] G. Heimeriks, M. Hörlesberger, P. Van den Besselaar, Mapping communication and collaboration in heterogeneous research networks, Scientometrics 58 (2) (2003) 391–413.

[142] I. Gómez, M. Bordons, M. Fernandez, A. Méndez, Coping with the problem of subject classification diversity, Scientometrics 35 (2) (1996) 223–235.