Measuring diversity in heterogeneous information networks

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, Fabien Tarissan

Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Measuring diversity in heterogeneous information networks

Pedro Ramaciotti Morales [a,b,*], Robin Lamarche-Perrin [c], Raphaël Fournier-S'niehotta [d], Rémy Poulain [b], Lionel Tabourier [b], Fabien Tarissan [e]

[a] *Sciences Po, médialab, Paris, France*
[b] *Sorbonne Université, CNRS, LIP6, F-75005 Paris, France*
[c] *CNRS, ISC-PIF, France*
[d] *CEDRIC, CNAM, Paris, France*
[e] *Université Paris-Saclay, CNRS, ISP, ENS Paris-Saclay, Cachan, France*

ARTICLE INFO

Article history:
Received 11 January 2020
Received in revised form 14 December 2020
Accepted 5 January 2021
Available online xxxx
Communicated by R. Klasing

Keywords:
Diversity measures
Heterogeneous information networks
Random walks on graphs
Recommender systems
Social network analysis

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.

Contents

* Corresponding author at: médialab, Sciences Po, 27 rue Saint-Guillaume - 75337 Paris cedex 07, France.
E-mail address: pedro.ramaciottimorales@sciencespo.fr (P. Ramaciotti Morales).

https://doi.org/10.1016/j.tcs.2021.01.013
0304-3975/

1. Introduction

Diversity is a concept of importance in several different domains of research, such as ecology [1], economy [2], public policy [2], information theory [3,4], 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* (how different types of entities are between them) [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-Hirschman 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 [16–18]. 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 [21–23].

Diversity measures can be computed over different types of data in a multitude of contexts. Access to data traces of different real phenomena has enabled for a tremendous 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 [25] and recommender systems [26], as well as in the artificial intelligence and representation learning communities [27–29].

Much of the success of these representations and their precursors – such as *multi-layer graphs* [30,31] – 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 [32–34] or for recommendation purposes [26,35,36]. While the application of diversity measurements to graph structures is not new [37,38], it is gaining widespread use in different communities [39], and in particular in the information retrieval and recommender systems communities [40]. Few studies have hinted at the application of *entropy* [10] (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 [41]. 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 [42,43]. 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. And in developing this formal framework, we also 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 existing 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 new 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 $D : I \times T \to \mathbb{R}^+$ 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 $p_\tau(t) = \frac{a_\tau(t)}{|\tau|}$. Using these definitions, we further narrow our consideration of diversity measures to functions that map proportional abundances resulting from a classification to non-negative real values: $D(\tau) = D(p_\tau(t_1), \ldots, p_\tau(t_k))$ with $k = |T|$ being the number of types. Hence, a diversity measure D is an application from Δ^* to $\mathbb{R}^+$, where $\Delta^* = \cup_{k \geq 0} \Delta^k$ is the union of all standard k-simplices, that is the set of probability distributions on discrete spaces of size $k+1$:

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

2.2. The diversity of diversity measures

As stated in the previous subsection, the term *diversity* is used to designate various properties of dissimilarity in a range of domains, such as ecology [44–46,1], life sciences [47], economics [2,12], public policy [48–50], information theory [3,4], internet & media studies [5,6], physics [51,52], social sciences [53], complexity sciences [54,55,7,8], and opinion dynamics [56]. 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 [57] for an extended discussion of these properties.

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

Richness [58,59] 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.

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

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 [60, Section 9].

Shannon entropy [10,61], 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:

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

Many classical diversity measures are functions of the properties identified as *variety* and *balance*. Shannon entropy, introduced in the context of channel capacity in telecommunications, is clearly affected by proportional abundances, and thus by their *balance*, but also by their *variety*: according to Shannon entropy, a bookcase with books that are uniformly distributed among 5 types is more diverse than a bookcase with books that are uniformly distributed among 4 types. By applying normalization, one may restrain the measurement to the property of *balance*. The diversity measure known as **Shannon Evenness** [62] in ecology, for example, consists of the ratio between measured entropy and maximal entropy for the same number of effective types and only accounts for the property of *balance*.

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 [63]. In deep learning methods for artificial intelligence, Shannon entropy is often used for quantifying information gain [64].

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:

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

Related to the *variety* and *balance* properties, this index (also known as the **Simpson Index** [65]), was first introduced by Hirschman [66] and later by Herfindahl [67] 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 [68] and **Population Heterozygosity** in genetics [69]) is another prominent example of a measure accounting for *variety* and *balance*. Also known as the probability of interspecific encounters in ecology [70], 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** [71], 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

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

Other diversity measures address only the property of *balance*. The **Berger-Parker Index** [72], 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:

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

It is easy to see that only the *balance* property affects this diversity measure. If the books of a 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 [73]. One example of a measure of *disparity* is proposed in [74]:

$$\mathrm{Disparity}(T) = \frac{1}{|T|\,(|T|-1)} \sum_{t,t' \in T} d(t,t'),$$

where d is a metric on the set T of types. *Disparity* is the underlying property in some use cases of the notion of diversity. Examples may be found in fields such as paleontology [75], economics [76], and biology [77]. Furthermore, diversity measures accounting for *disparity* as well as *variety* and *balance* exist [78].

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. [79–82]). Drawing from these existing axiomatizations, we propose an organization of axioms suited for the purposes of this article.

We first introduce four axioms that encode properties which are necessary for a diversity measure, *i.e.*, *symmetry*, *expansibility*, *transferability*, and *normalization*. Then, we present a family of functions that satisfy these properties. By imposing an additional property known as *replicability* in the form of an axiom, the resulting measures of the theory correspond to the family of functions known as *true diversities*. One 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 $D : \Delta^{k-1} \to \mathbb{R}^+$, a probability distribution $p = (p_1, \ldots, p_k) \in \Delta^{k-1}$, 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 σ 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 measures 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 does not contain any books does not change the diversity of a bookcase.

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

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

For a diversity measure to be a measure of balance it needs to satisfy the *transfer principle*, also called the *Pigou-Dalton principle* [83]: 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

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

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

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

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

Proof. By the application of Axiom 2 and Axiom 1, the claim of the theorem is equivalent to

$$D(\underbrace{\ldots, p_i, p_{i+1}, \ldots}_{k \text{ entries}}) \quad \geq \quad D(\underbrace{\ldots, p_i + p_{i+1}, 0, \ldots}_{k \text{ entries}}).$$

Without loss of generality, let us suppose that $p_i \geq p_{i+1}$, and let us apply the transfer principle of Axiom 3 to this distribution $(\ldots, p_i + p_{i+1}, 0, \ldots)$ of k entries. The first condition for its application is always satisfied, *i.e.*, $p_i + p_{i+1} > 0$ (if $p_i + p_{i+1} = 0$ the theorem is trivially assured by Axioms 2 & 1). Choosing $\epsilon = p_{i+1}$ satisfies the second condition of application, because $p_{i+1} \leq (p_i + p_{i+1} - 0)/2$ if $p_i \geq p_{i+1}$. Finally, the application of Axiom 3 gives the desired result. ■

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

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

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

Proof. The second inequality is warranted by Theorem 1. If distribution $p = (p_1, \ldots, p_k) \in \Delta^{k-1}$ is the uniform distribution, that we shall denote by u, the first inequality is trivially satisfied. If, on the other hand, p is any distribution that is not uniform, we will show that Axiom 3 assures the construction of a sequence of m distributions $p^1, \ldots, p^m$ in Δ^{k-1}, such that $p^0 = p$, p^m is the uniform distribution, and $D(p^1) \leq D(p^2) \leq \ldots \leq D(p^m)$, thus assuring that $D(u) \geq D(p)$. We do this by adapting the proof of [14, Thm. 1] developed for measures of concentration. Because of Axiom 1, we can set, without loss of generality, p^1 as the distribution that results from ordering the entries of p in decreasing order, still resulting in $D(p^1) = D(p)$. Given a non-uniform distribution p^l of the sequence, and assuming that its entries are arranged in decreasing order, we will show how to compute the next distribution of the sequence, p^{l+1}, so that $D\left(p^{l+1}\right) \geq D\left(p^l\right)$, using Axiom 3. Let δ^l be a vector in $\mathbb{R}^k$ resulting from subtracting p^l and u element-wise: $\delta^l_i = p^l_i - u_i = p^l_i - 1/k$. Now let i^- be the first negative entry of δ^l: $i^- = \min\left\{1 \leq i \leq k : \delta^l_i < 0\right\}$. Because entries of p^l cannot all be less than $1/k$, we know that i^- is never the first entry ($i^- > 1$), and because entries cannot all be greater than $1/k$, we know that i^- can always be

determined ($1 < i^- \leq k$), as long as p^l is non-uniform. Because p^l is not uniform, we know that $p^l_1 > 1/k$, and so we transfer the quantity $\min(\delta^l_1, -\delta^l_{i^-})$ from entry p^l_1 to entry p^- to compute a distribution $\overline{p}^{l+1}$. The components of $\overline{p}^{l+1}$ are computed as: $\overline{p}^{l+1}_1 = p^l_1 - \min(\delta^l_1, -\delta^l_{i^-})$, $\overline{p}^{l+1}_{i^-} = p^l_{i^-} + \min(\delta^l_1, -\delta^l_{i^-})$, and $\overline{p}^{l+1}_i = p^l_i$ for $i \notin \{1, i^-\}$. Next, we compute p^{l+1} as the distribution resulting from arranging the elements of $\overline{p}^{l+1}$ in decreasing order. Because either entry $i = 1$ or entry $i = i^-$ was set to $1/k$, a new entry is now $1/k$ (as entries in u). And because, at each step in the sequence, a new entry is set to $1/k$, we know that this sequence is finite. ■

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

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

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 [84] is the family of *quasilinear means* (developed by Kolmogorov [85] and Nagumo [86]). Quasilinear means of a probability distribution are central to the quantification of information in information theory [3], and are of the form

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

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 ϕ a strictly monotonic increasing continuous function.

A sub-family of quasilinear means, the so-called self-weighted quasilinear means [87], has additional properties that will be of interest in what follows.

Definition 1 *(Self-weighted quasilinear means [87]).* A function $S : \Delta^* \to \mathbb{R}^+$ is a self-weighted quasilinear mean if it is of the form

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

with ϕ a strictly monotonic increasing continuous function.

Further restrictions of the considered diversity measures, described by the following theorem, result in a family of functions that simultaneously satisfy the above properties described by the axioms.

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

Proof. Let us consider a diversity measure D in the form of the reciprocal of a self-weighted quasilinear mean: $D(p) = \frac{1}{S(p)}$, with S of the form given in Definition 1, with ϕ continuous strictly increasing such that $h(t) = t\phi(t)$ is concave. It is easy to check that D satisfies Axiom 1 (symmetry) because of the commutativity of the sum. Because the summands are self-weighted, adding new zero-valued entries results in zero-valued summands, which assures that Axiom 2 (expansibility) is

satisfied. By construction, uniform distributions of k entries have diversity $\frac{1}{\phi^{-1}(k\cdot(1/k)\cdot\phi(1/k))} = k$, assuring that D satisfies Axiom 4 (normalization).

Finally, given $p = (\dots, p_i, \dots, p_j, \dots)$ with $p_i > p_j$ and $\epsilon \leq (p_i - p_j)/2$, let us consider $\tilde{p} = (\dots, p_i - \epsilon, \dots, p_j + \epsilon, \dots)$. If $S(\tilde{p}) \leq S(p)$, then $D(\tilde{p}) \geq D(p)$ and D would satisfy Axiom 3 (transfer principle). Because ϕ is monotonic strictly increasing, ϕ^{-1} also is, and $S(\tilde{p}) \leq S(p)$ if $\sum_{l=1}^{k} h(p_l) \geq \sum_{l=1}^{k} h(\tilde{p}_l)$. Because p and $\tilde{p}$ share all but the i-th and j-th entries, this last inequality is assured by

$$h(p_i) - h(p_i - \epsilon) + h(p_j) - h(p_j + \epsilon) \geq 0.$$

To see that this inequality holds, let us note that we can always compute $\theta = (p_i - p_j - \epsilon)/(p_i - p_j - 2\epsilon)$, with $\theta \in (0, 1)$, so that $p_i = \theta(p_i - \epsilon) + (1 - \theta)(p_j + \epsilon)$ and $p_j = (1 - \theta)(p_i - \epsilon) + \theta(p_j + \epsilon)$. This step is adapted from the proof of [14, Thm. 3]. θ is always positive by the restrictions on ϵ required by Axiom 3. By concavity of h we can write inequalities for $h(p_i)$ and $h(p_j)$:

$$h(p_i) = h\left(\theta(p_i - \epsilon) + (1 - \theta)(p_j + \epsilon)\right) \geq \theta h(p_i - \epsilon) + (1 - \theta)h(p_j + \epsilon),$$
$$h(p_j) = h\left((1 - \theta)(p_i - \epsilon) + \theta(p_j + \epsilon)\right) \geq (1 - \theta)h(p_i - \epsilon) + \theta h(p_j + \epsilon).$$

Additioning these two inequalities we obtain $h(p_i) - h(p_i - \epsilon) + h(p_j) - h(p_j + \epsilon) \geq 0$. ■

Theorem 4 provides us with an explicit expression for functions that satisfy Axioms 1, 2, 3 & 4. The use of self-weighted quasilinear means yields, however, a subset of the functions defined by these axioms. Indeed, there are diversity measures that satisfy these axioms but cannot be expressed as self-weighted quasilinear means (*e.g.*, Hall-Tideman Index [88]).

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.

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

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

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

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

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

True diversities were first introduced as the Hill Number [13] and named *true diversity* in [89]. Variants of true diversities exist in different domains. The Hannah-Kay concentration index of order α [90] is the reciprocal of D_α. In information theory, Rényi Entropy [4] of order α, denoted by H_α, is the natural logarithm of D_α: $H_\alpha(p) = \ln D_\alpha(p)$.

Theorem 5 *(Diversity measures that satisfy the replication principle are true diversities [15]). Suppose that a diversity measure D can be represented as a reciprocal self-weighted quasilinear mean as in Theorem 4, then D is a true diversity for some order α if and only if D satisfies the replication principle of Axiom 5.*

The reader is referred to [15, Theorem 3.1] for the proof.

The replication principle may be needed to avoid otherwise paradoxical results in many applications. Let us consider for example a library with 3 bookcases, each containing items of 3 different types: 9 types of items organized in 3 bookcases, with no types dispersed in multiple bookcases. Let us also suppose that on each one of the 3 bookcases, the distribution of items into the 3 types is the same: 10-20-70%, *i.e.*, $p = (0.1, 0.2, 0.7)$ for each bookcase and

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

Order (α)	Name	True diversity	Expression	Relation to other diversity measures
0	Richness diversity	$D_0(p)$	$\lvert\{i \in \{1,\dots,k\} : p_i > 0 \in\}\rvert$	Same as richness [58,59].
1	Shannon diversity	$D_1(p)$	$\left(\prod_{\substack{i=1 \\ p_i \neq 0}}^{k} p_i^{p_i}\right)^{-1}$	Exponential of Shannon entropy [10,61]: $H(p) = \log_2(D_1(p))$, with H in base 2.
2	Herfindahl diversity	$D_2(p)$	$\left(\sum_{i=1}^{k} p_i^2\right)^{-1}$	Reciprocal of the Herfindahl-Hirschman Index [12]: $\mathrm{HHI}(p) = 1/D_2(p)$.
∞	Berger diversity	$D_\infty(p)$	$\left(\max_{i\in\{1,\dots,k\}} \{p_i\}\right)^{-1}$	Reciprocal of the Berger-Parker Index [72]: $\mathrm{BPI}(p) = 1/D_\infty(p)$.

$p = \left(\frac{0.1}{3}, \frac{0.2}{3}, \frac{0.7}{3}, \frac{0.1}{3}, \frac{0.2}{3}, \frac{0.7}{3}, \frac{0.1}{3}, \frac{0.2}{3}, \frac{0.7}{3}\right)$ for the library of 3 bookcases. Finally, let us now suppose that, due to maintenance costs, 2 of our 3 bookcases will have to be discarded, and that we are interested in measuring the diversity that will be lost, and the diversity we will manage to preserve. If we consider the Gini-Simpson Index (cf. Section 2.2), we measure the initial diversity of our 3 bookcases at 0.82, the diversity of the saved bookcase at 0.46, and the diversity of the 2 lost bookcases at 0.73. Paradoxically, because the Gini-Simpson Index does not satisfy the replication principle, we have saved about 56.1% ($\frac{0.46}{0.82}$) of the initial diversity, but we have lost about 89% ($\frac{0.73}{0.82}$) of it. Had we taken the true diversity of order 1 for measurements, the initial diversity would have been of 6.69, while that of the saved bookcase would have been 2.23, and that of the 2 lost bookcases would have been 4.46. Because true diversities satisfy the replication principle, this would have yielded no paradox: we would have measured a lost of 2/3 of the diversity while measuring 1/3 of the initial diversity saved. The replication principle allows for interesting algebraic properties of diversity measures: when gathering or disassembling multiple distributions, this principle ensures that sum of diversities is preserved. For further examples, and for a discussion of the implications of the replication principle in ecology, the reader is referred to [91,92].

True diversities are related to several of the diversities used in different domains and identified in Section 2.2. *Richness* of a distribution p can be computed as the limit of $D_\alpha(p)$ when $\alpha \to 0^+$, observing that $p_i^\alpha \to 1$ if $p_i > 0$, thus resulting in the count of effective types. We thus identify richness with 0-order true diversity, calling it **Richness diversity**. $D_1(p)$, 1-order true diversity (or **Shannon diversity**), also called *perplexity* [93], 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: $D_2(p) = 1/\mathrm{HHI}(p)$. The Berger-Parker Index is also identified with the result of a limit process. By observing that $D_\alpha \xrightarrow{\alpha\to\infty} 1/\max\{p_1,\dots,p_k\}$ (Section 5.4 of [3]) we can define

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

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

Let us illustrate some of these properties in Fig. 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 Fig. 1). However, when the distribution into types is not uniform, these measures behave differently (decreasing curves in Fig. 1). In this case, parameter α expresses the way non-uniformity, or *balance*, is taken into account. If α 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 α is high, inequalities in a distribution will have a strong impact on 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 Fig. 1 illustrate how parameter α can modulate the relative importance given to *variety* and *balance* (cf. Section 2.2): a distribution with 6 types could be evaluated less diverse than one with 4 types if it is sufficiently unbalanced for a given value of α. 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.

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

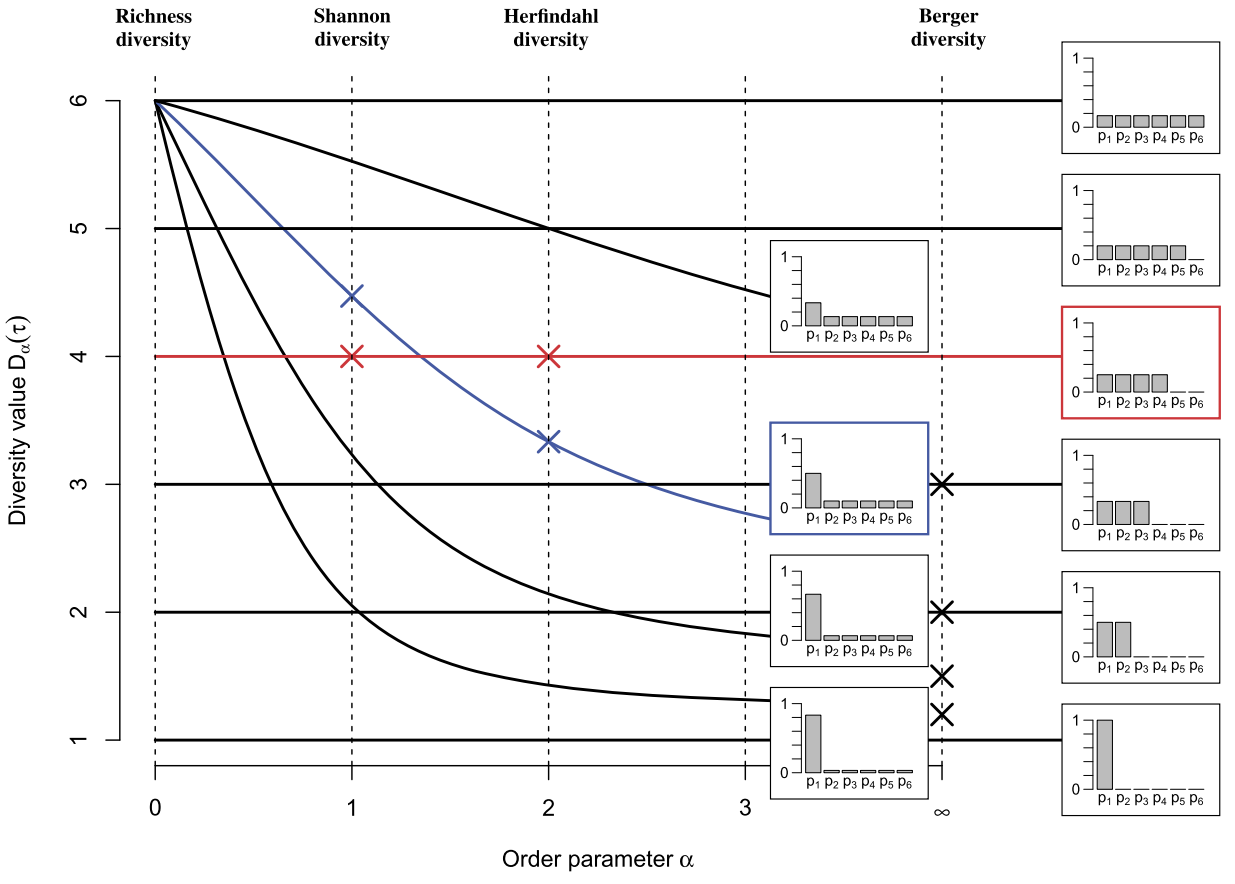


Fig. 1. Values of different true diversities, depending on order α, for different distributions. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

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* [4,94]. Among these generalizations, a well-known special case is the generalization of *Shannon entropy* to *Kullback-Leibler divergence* (also known as *relative entropy*) [95,96].

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

Definition 3 *(Relative true diversity).* The relative true diversity of order α is 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 $\alpha \in \mathbb{R}^+$,

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

As with true diversities, extreme values are defined as the result of limit processes (cf. Theorems 4, 5, & 6 of [94]):

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

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

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

$D_\alpha(p \parallel u) = k/D_\alpha(p)$, and thus $D_\alpha(p \parallel u) \in [1, k]$ (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.*, comics, novels) but also according to their author.

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

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

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

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

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

The first of two additivity principles 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 τ_1 and τ_2 such that $p_{\tau_1 \times \tau_2}(t_1, t_2) = p_{\tau_1}(t_1) p_{\tau_2}(t_2)$, we have $D\left(p_{\tau_1 \times \tau_2}\right) = D\left(p_{\tau_1}\right) D\left(p_{\tau_2}\right)$.

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 [80]). True diversities D_α 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 τ_1 and τ_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 τ_1 and τ_2, we have $D\left(p_{\tau_1 \times \tau_2}\right) = D\left(p_{\tau_1}\right) D\left(p_{\tau_2 \mid \tau_1}\right)$ where $D\left(p_{\tau_2 \mid \tau_1}\right) = \prod_{t_1 \in T_1} D\left(p_{\tau_2 \mid t_1}\right)^{p_{\tau_1}(t_1)}$.

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 [80]). 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 in [96]): $H(X, Y) = H(X) + H(Y|X)$ for random variables X and Y.

Theorem 7 will justify the use of 1-order diversities in some results regarding the relations of different *network diversity measures* in the next sections. Fig. 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.

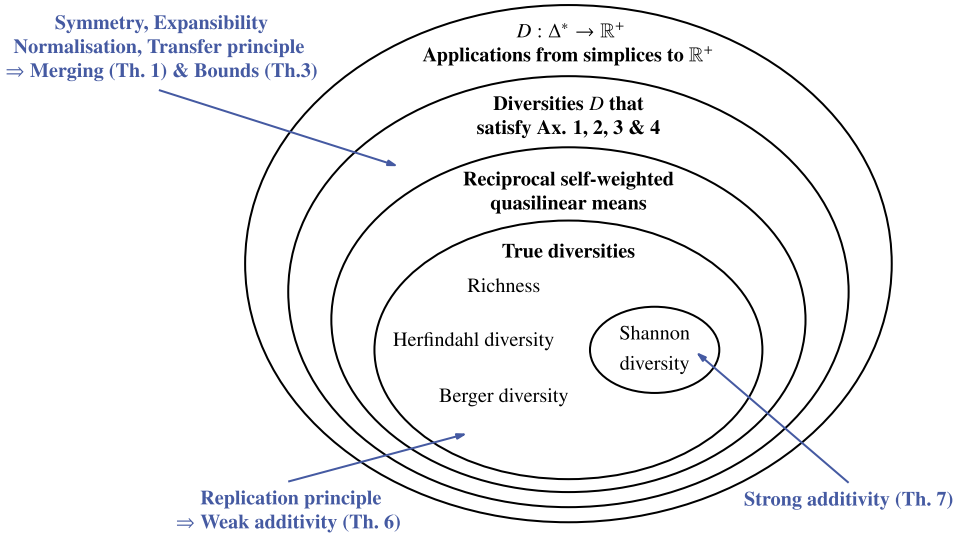


Fig. 2. Relations between the different families of diversity measures, and their most important properties.

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 [97–99]). 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 5, many situations encountered in practice can be represented using heterogeneous information networks. For example, when modeling the consumption of news on a website, the situation may be represented as users selecting articles, and 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

We consider a multigraph composed of a set of nodes $\boldsymbol{V}$, linked by a set of directed edges $\boldsymbol{E}$. We propose the following system of capitalization and typefaces to 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 notated 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 $\mathcal{V}$ and $\mathcal{E}$;
- random variables with support on sets of vertices are notated by the capital letter X.

3.2. Heterogeneous information networks

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

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

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

We now define heterogeneous information networks using multigraphs. Classical heterogeneous information networks can be easily accounted for by constraining the multiplicity of edges.

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

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

Label functions φ and ψ, 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 $\mathcal{R} = \{R_1, \ldots, R_M\}$, φ and ψ induce partitions $\mathcal{V} = \{V_1, \ldots, V_N\}$ on $\boldsymbol{V}$ and $\mathcal{E} = \{E_1, \ldots, E_M\}$ on $\boldsymbol{E}$. These partitions are such that $\forall v \in \boldsymbol{V}$, $(\varphi(v) = A_i \Leftrightarrow v \in V_i)$ and $\forall e \in \boldsymbol{V}$, $(\varphi(e) = R_j \Leftrightarrow e \in E_j)$. Thus, abusing notation, we make indistinct use of types in $\mathcal{A}$ and sets in $\mathcal{V}$, and of types in $\mathcal{R}$ and sets in $\mathcal{E}$ when this is not ambiguous.

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

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

As before, we also define:

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

Following the example of existing definitions for heterogeneous information networks [24,25,100], 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. Fig. 3 illustrates a heterogeneous information network and its network schema.

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

Knowledge graphs (*e.g.*, Google's Knowledge Graph [101]) are knowledge-based systems closely related to heterogeneous information networks. They are used to store complex structured and unstructured data in the form of a network, based on the Resource Description Framework (RDF) [102], which models data as entries of the form ⟨*Subject, Property, Object*⟩. If edges of a same type always link source vertices of the same type with target vertices of the same type (cf. Definition 6), it is easy to see that identifying a *Property* in the RDF data model with an edge type allows for the identification of a knowledge graph with a heterogeneous information network [100]. Early pairings of the two concepts were proposed to leverage the heterogeneous information network formalism in data mining tasks in knowledge graphs [103]. While some works have equated these two closely similar concepts [104], most insist in differentiating heterogeneous information networks as a mathematical formalism suitable for the treatment of data mining problems using knowledge graph data [105,106].

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 $V_{\mathrm{src}}(E)$ is connected to at least one vertex in $V_{\mathrm{dst}}(E)$, *i.e.*, $\forall v_1 \in V_{\mathrm{src}}(E)$ $(\epsilon_E(v_1, -) > 0)$, we denote by $p_E : V_{\mathrm{src}}(E) \times V_{\mathrm{dst}}(E) \to [0, 1]$ the *transition probability* of the random walk following edges in E, for all $(v_1, v_2) \in V_{\mathrm{src}}(E) \times V_{\mathrm{dst}}(E)$, as

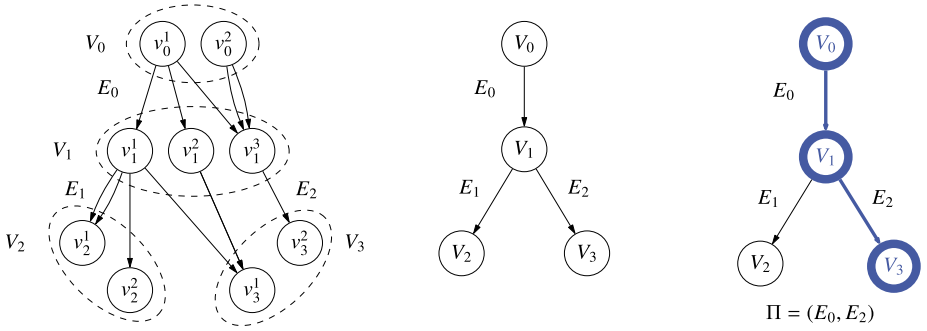


Fig. 3. A heterogeneous information network (left), its network schema (center), and a meta path Π on the network schema (right).

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

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

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

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

3.3. Meta paths and constrained random walks

Random walks in heterogeneous information networks can be constrained [33,107] to follow a specific sequence of edge types, called *meta path* [100,108]. This enables 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 we propose in the next section.

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

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

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

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

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

This is known as a *path-constrained random walk* in the information retrieval community [33,107]. It follows from Definitions 8 and 11 that a random walk restricted to a meta path Π of length k is a Markov chain with transition probabilities defined as

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

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

For the next two definitions, we consider a meta path $\Pi = (E_{r_1}, \ldots, E_{r_k})$ of length k and its associated random walk restricted to Π, *i.e.*, the sequence $(X_0, X_1, \ldots, X_k)$ of random variables. The probability distribution in $V_{\mathrm{dst}}(\Pi)$ 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 $X_k \in V_{\mathrm{dst}}(\Pi)$, that is, the destination vertex of the random walk constrained to Π, given that it started in $v_0 \in V_{\mathrm{src}}(\Pi)$ (*i.e.*, $X_0 = v_0$), is denoted by $p_\Pi(v_k \mid v_0)$ for $v_k \in V_{\mathrm{dst}}(\Pi)$ and can be recursively computed as follows:

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

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

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

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

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

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

We now consider the edges resulting from the projection of all edge types in a meta path Π. This operation, related to the counting of paths in meta paths, is used in the literature in related measures, such as the construction of similarity metrics for vertex searches [32] or for recommender systems [26].

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

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

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

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

4. Network diversity measures

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

Depending on the chosen meta 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

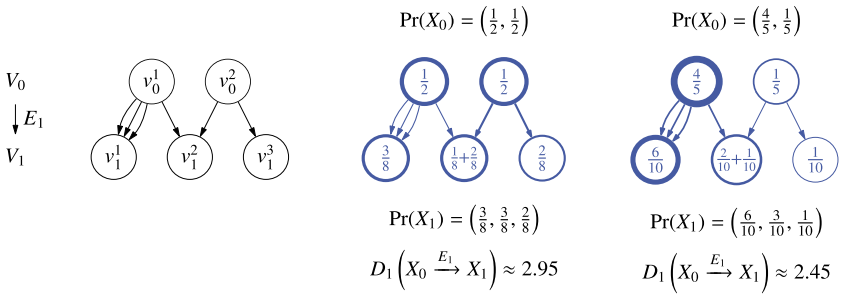


Fig. 4. Computation of the collective V_1 diversity of V_0 along a simple meta path made of only one edge type. Diversity along a path type depends on the starting probability distribution $\Pr(X_0)$ and on transition probabilities.

associated with meta paths may correspond to different concepts will be made clear in this section, and illustrated through different applications in the next section.

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

4.1. Collective and individual diversities

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

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

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

Note that this measure depends on the starting probability distribution $\Pr(X_0 = v_0)$ and on transition probabilities $p_{E_{r_i}}(v_i \mid v_{i-1})$ for each $E_{r_i} \in \Pi$. Fig. 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 Fig. 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 $X_0 \sim \text{Uniform}(V_{\text{start}})$ 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 $V'_{\text{start}} \subset V_{\text{start}}$ and choosing $X_0 \sim \text{Uniform}(V'_{\text{start}})$ allows us to define the collective diversity of the group of nodes V'_{start}.

Conditioned probabilities of random walks along a path Π are also of interest, as they convey information about the structure of the network reachable from some vertices in V_{start}. In particular, given a starting vertex $v_0 \in V_{\text{start}}$, we define the *individual V_{end} diversity of v_0 along* Π as the true diversity of the probability distribution of $X_k \in V_{\text{end}}$ at the end of the constrained random walk, knowing that it started at a given vertex $v_0 \in V_{\text{start}}$ (*i.e.*, $X_0 = v_0$). Fig. 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 $v_0 \in V_{\text{start}}$, we define the *individual V_{end} diversity of v_0 along* Π as the true diversity of the probability distribution of the ending vertex of the constrained random walk. We denote it as $D_\alpha\left(X_0 \xrightarrow{\Pi} X_k \mid X_0 = v_0\right)$ and compute it as follows:

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

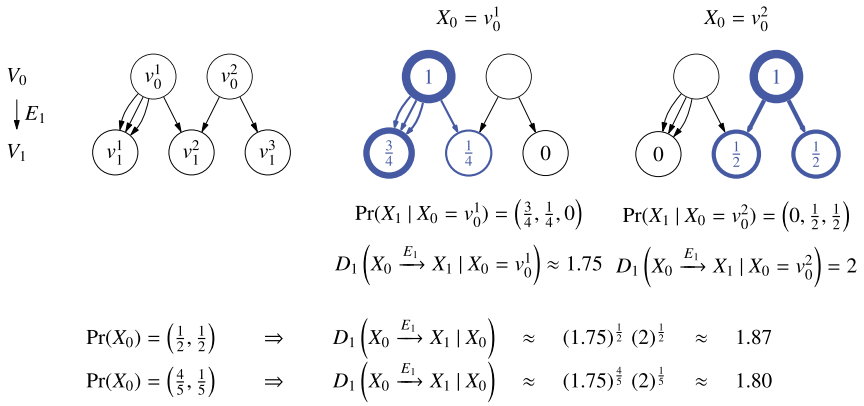


Fig. 5. Examples of a heterogeneous information network (top left), the individual diversities of two vertices (top center and top right), and the mean individual diversities for two different starting probability distributions (bottom).

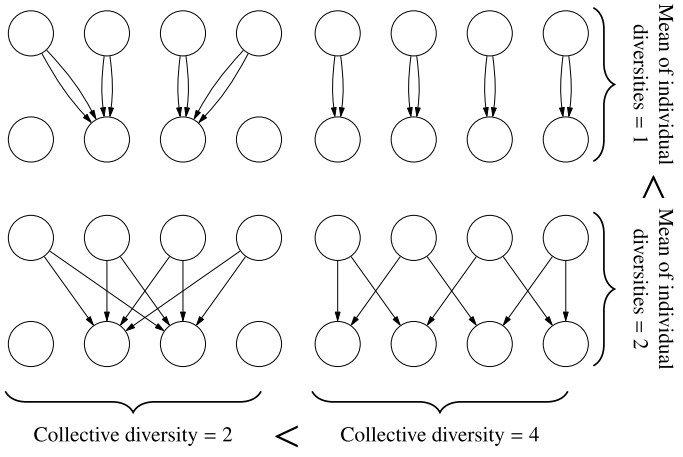


Fig. 6. Different heterogeneous information networks with two vertex types, illustrating different relative ordering for collective and mean individual diversities.

An aggregation of individual diversities may 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 [96]), we define the *mean V_{end} individual diversity of V_{start} along* Π as the weighted geometric mean of individual diversities.

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

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

This mean is weighted by –and so depends on– the starting probability distribution $\Pr(X_0)$ over V_0, and 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 $X_0 \sim \text{Uniform}(V_{\text{start}})$). As with collective diversity, the mean individual diversity of a vertex group $V'_{\text{start}} \subset V_{\text{start}}$ can be considered choosing $X_0 \sim \text{Uniform}(V'_{\text{start}})$. Fig. 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 Fig. 6. It is possible for a system to have a low mean individual diversity while having a high collective diversity (top-right in Fig. 6), or a high mean individual diversity while having a low collective diversity (bottom-left in Fig. 6).

4.2. Backward diversity

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Relative diversities (to be illustrated in Section 5) are useful whenever we want to compare the diversity related to some meta path with a baseline resulting with another one. 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_Π of a meta path Π (cf. Definition 14 in Section 3.3), we can also define the diversity of the distribution on V_{end} of a constrained random walk starting at $v_0 \in V_{\mathrm{start}}$ and following the edges in E_Π.

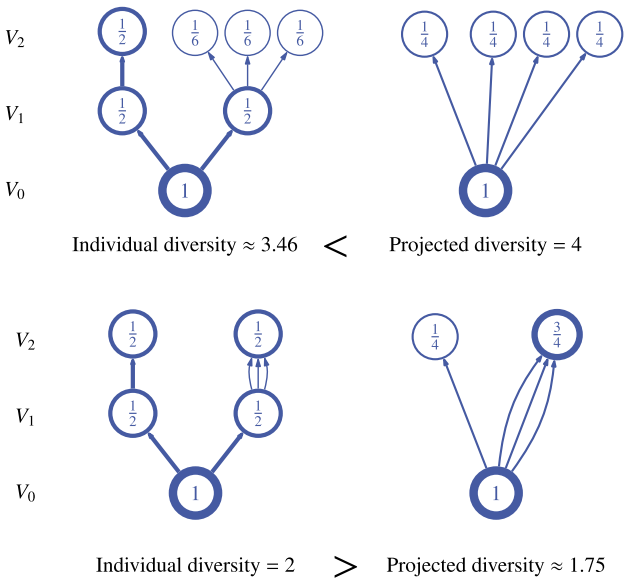


Fig. 7. Individual Shannon diversities of a meta path on two heterogeneous information networks, compared with the resulting projected diversities. We illustrate two situations: one in which projected diversity is greater than individual diversity (top), and one where individual diversity is greater than projected diversity (bottom).

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

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

Note that in the previous definition, $p_{E_\Pi|v_0}$ is the probability distribution from Definition 12 when the meta path is made only of projected edges in E_Π. Fig. 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 $1 \le i \le j \le k$, its restriction

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

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

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

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

In other words, 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 an only if all individual backward diversities along this remaining part are minimal.

Before proving Theorem 8, let us first prove the following result linking 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:*

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

where $D_1(\Pr(X_0))$, 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

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

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

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

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

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

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

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

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

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

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

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

A first application of the strong additivity principle yields

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

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

A second application of the strong additivity principle yields

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

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

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

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

This gives the desired result

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

from which it follows that

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

if mean backward diversity is not equal to 1. ■

4.6. Summary of network diversity measures

In this Section 4, 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 different 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, 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 numerical example of application of the network diversity measures to real datasets in Section 5.2. In a third example (Section 5.3), we provide a detailed application case in a recommender system setting, 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 5.4. Finally, we provide other examples of applications in ecology in Section 5.5, antitrust regulation in Section 5.6, and scientometrics in Section 5.7.

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 3.1, for the sake of readability, let us denote these vertex types respectively

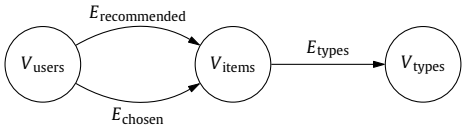


Fig. 8. Network schema of a simple heterogeneous information network, where users in V_{users} have chosen and have been recommended items in V_{items}, which are classified into types in V_{types}.

by V_{users}, V_{items}, and V_{types}. An example of entities represented by items is a set of films, and an example of types is then a set of film genres (e.g., comedy, thriller).

Let us now consider three edge types, indicating items chosen by users, items recommended to users, and classification of items into types. We respectively denote these edge types as E_{chosen}, $E_{\text{recommended}}$, and E_{types}. Fig. 8 illustrates the network schema of the described heterogeneous information network.

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

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

which 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):

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

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

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

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

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

5.2. A numerical example

We turn now to an empirical example that will show how our network diversity measures can be used and interpreted in order to analyze a specific dataset. The following study deals with the behavior of users on online musical platforms. In

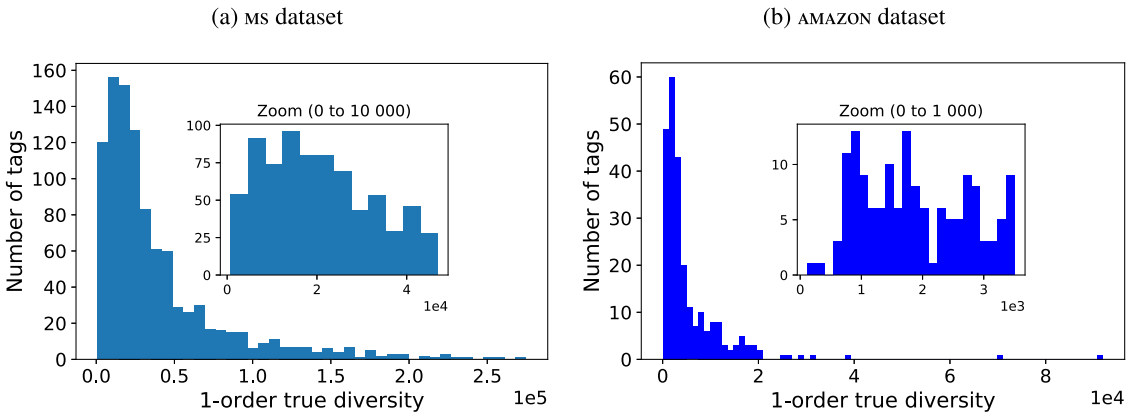


Fig. 9. Histograms showing the distributions of the 1-order true diversities of the tag audience ($D_1(p_{\Pi_{\text{audience}}|t})$ for all tags in V_{tags}) for the MS (left) and the AMAZON (right) datasets.

such platforms, users listen to songs and songs are usually tagged by musical categories. This situation can be modeled by a heterogeneous information network with three vertex types (users V_{users}, songs V_{songs}, and tags V_{tags}) and two types of edges (E_{consumed} connecting V_{users} to songs V_{songs}, and E_{tagged} connecting V_{songs} to tags V_{tags}). We use network diversity measures to investigate two different questions. First we analyze the diversity of the distribution of users that listen to songs tagged with a given tag $t \in V_{\text{tags}}$: the *diversity of the tag audience*. Second, we analyze, for a given user $u \in V_{\text{users}}$, the diversity of the distribution of tags associated with the songs she listens to: the *diversity of a user's attention*. Translated into our framework, we consider the meta-paths $\Pi_{\text{audience}} = (E^{\intercal}_{\text{tagged}}, E^{\intercal}_{\text{consumed}})$ and $\Pi_{\text{attention}} = (E_{\text{consumed}}, E_{\text{tagged}})$ to analyze the following diversities:

- $\forall t \in V_{\text{tags}}$, $D_\alpha(p_{\Pi_{\text{audience}}|t})$: the individual diversities of audiences of tags in V_{tags} (see Section 5.2.2);
- $\forall u \in V_{\text{users}}$, $D_\alpha(p_{\Pi_{\text{attention}}|u})$: the individual diversities of tags of users in V_{users} (see Section 5.2.3).

It is worth noting that while the numerical analyses presented in this section are new, a complete study dedicated to this context and dataset has been published in [109], which can provide a useful complement to the reader.

5.2.1. Datasets used

In this numerical example we use the same above-described network schema (cf. Definition 7) for V_{users}, V_{songs}, and V_{tags} to analyze data from two different datasets that can be modeled by this heterogeneous information network.

For the first dataset, we use data from the *Million Song Dataset* project [110]. In particular, we use the *user-taste-profile* data,[1] that contains 48 million events of users listening to songs, to determine V_{users}, V_{songs}, and E_{consumed} parts of a heterogeneous information network model. For the V_{tags} and E_{tagged} parts, we use data from the *last.fm* dataset[2] that provides a list of tags for each song. Using these two sources of data resulted in a dataset with 1,019,190 users in vertex type V_{users}, 234,379 songs in vertex type V_{songs}, and 1,000 tags in vertex type V_{tags}. We refer to this dataset as the MS data.

For the second dataset, we consider a collection of reviews made on *Amazon* [111,112], and that contain musical items (*e.g.*, CDs, vinyls, and digital music). From these data, we only retain the link between a user and a product (a song or an album here). *Amazon* provides a hierarchy of categories for each product, which allows us to extract musical tags for each song. This resulted in a dataset with 465,248 users in vertex type V_{users}, 445,514 songs in vertex type V_{songs}, and 250 tags in vertex type V_{tags}. We will refer to this second dataset as AMAZON.

Using the specified heterogeneous information network schema to model the data from these two datasets, we may use our network diversity measures to compute diversities in the data. For ease of analysis, we restrain our analysis in this Section 5.2 to the 1-order true diversity, *i.e.*, the Shannon diversity (cf. Table 1). The reader is referred to [109] for a study of these datasets using different diversity measures.

5.2.2. 1-order true diversity of the tag audience

First we focus on the diversity of the audiences of tags: the individual user-diversities of the tags. Fig. 9 presents the distribution of the 1-order true diversities $D_1(p_{\Pi_{\text{audience}}|t})$ of all tags $t \in V_{\text{tags}}$. We compute and present these values using the MS (Fig. 9a) and the AMAZON (Fig. 9b) datasets.

Both plots show strongly heterogeneous distributions of individual diversities: if most of the tags have a rather narrow audience, one can identify some tags with a particularly high diversity. This is the case for the tags *Rock* and *Pop* in both

[1] Available at https://labrosa.ee.columbia.edu/millionsong/tasteprofile.
[2] Available at https://labrosa.ee.columbia.edu/millionsong/lastfm.

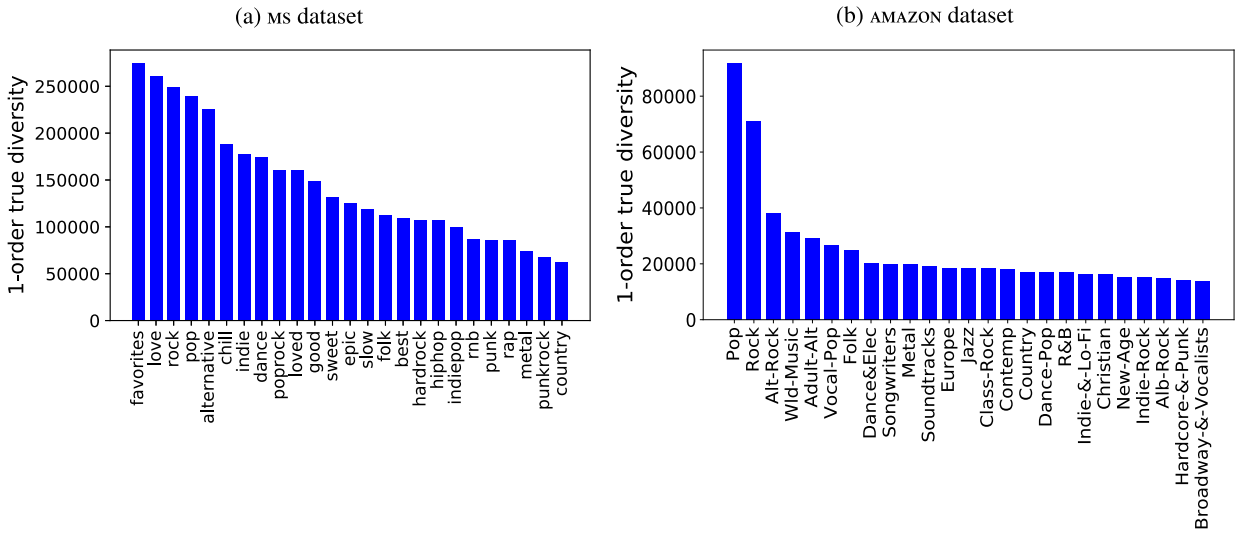


Fig. 10. Ordered 1-order true diversities of the tag audience $D_1(p_{\Pi_{\text{audience}}|t})$ for 25 selected tags for the MS (left) and the AMAZON (right) datasets.

datasets (see Fig. 10). But even for those tags, their diversity value (around 10^5 in MS and 10^4 in AMAZON) is still one order of magnitude lower than the maximal theoretical values: 1,019,190 for MS and 465,249 for AMAZON (cf. Axiom 4). One can, however, nuance this observation by noticing that small diversity values are more homogeneously distributed. This is visible in the insets of Fig. 9, which focus on the distribution of diversity values that are lower than 10,000 (74% of the nodes in MS, Fig. 9a) and lower than 1,000 (56% of the nodes in AMAZON, Fig. 9b). One can see in particular that the values are well distributed around the mean value of the dataset (respectively 24,850 for MS and 2,905 for AMAZON). This indicates that while one may spot some extremely diverse musical contents (*Rock* and *Pop*, for instance), most of them are narrowed towards a smaller and less diverse set of users (such as *Country* and *Punkrock* in MS or *New-Age* in AMAZON).

In order to further investigate how such diversity measures can be used to analyze specific categories, we show in Fig. 10 a selection of 25 tags for the two datasets. It is worth mentioning here that for the MS dataset, the tags are actually provided by the users themselves that can decide to use any word to tag any song (this is known as a *folksonomy* [113]). While most tags coincide with common music genres (like *Rock*, *Pop*, *Folk*, *Metal*, ...), others are obviously meant to give an appreciation of the songs (like *Favorites*, *Love*, *Best*, ...) or even to depict a moment at which a song is listened to (like *BeforeSleep* or *InShower*). The wide range of usage of the tags is an opportunity for us to assess how our network diversity measures respond to those different behaviors. For instance, one can expect tags like *Favorites* to be related to a broader and more diverse audience than *Metal* since the songs tagged by the former do not belong to a dedicated musical category. This is indeed confirmed by Fig. 10a which shows that popular tags like *Favorites* and *Love* have a diversity higher than any other tags of the dataset.

In contrast with the case of the MS dataset, the classification imposed by *Amazon* provides only tags that describe musical genres. This allows for a direct comparison of the musical categories presented in Fig. 10b, which provides interesting insights on the way users commit to the different categories. For instance, if we compare *Adult-Alternative* and *World-Music* with *R&B* and *Dance-Pop*,[3] it is remarkable that the two former ones have a diversity twice higher although the four tags have songs reviewed by the same number of users (approximately 300,000 users). This is a clear indication that users posting reviews on *Adult-Alternative* and *World-Music* songs are much more committed (the reviews are more uniformly distributed among the users) than the ones of *R&B* and *Dance-Pop*.

5.2.3. 1-order true diversity of users' attention

We now turn to the diversity of users' attention, the diversity of tags listened by users. Fig. 11 presents the distribution of the 1-order true diversities $D_1(p_{\Pi_{\text{audience}}|t})$ for all users $u \in V_{\text{users}}$. We compute and present these values using the MS (Fig. 11a) and AMAZON (Fig. 11b) datasets. In contrast with the distributions presented in Fig. 9, the diversity of users' attention is clearly homogeneous and centered around small values (compared to the maximal theoretical ones). This indicates that even if some users have a particularly high diversity, the vast majority of them have a relatively narrow consumption of the musical products. It is worth noting that, compared to the study of tags, that often had a meaningful name, we have no information regarding the profile of a user who is just an anonymized value in the dataset. Thus we cannot focus on specific users to provide an interpretation of the diversity values like we did in the previous section.

[3] We discard in the comparison *Rock* and *Pop* that have a particularly large number of users posting reviews to their songs, at least ten times higher than the number of users for any other tag in the dataset.

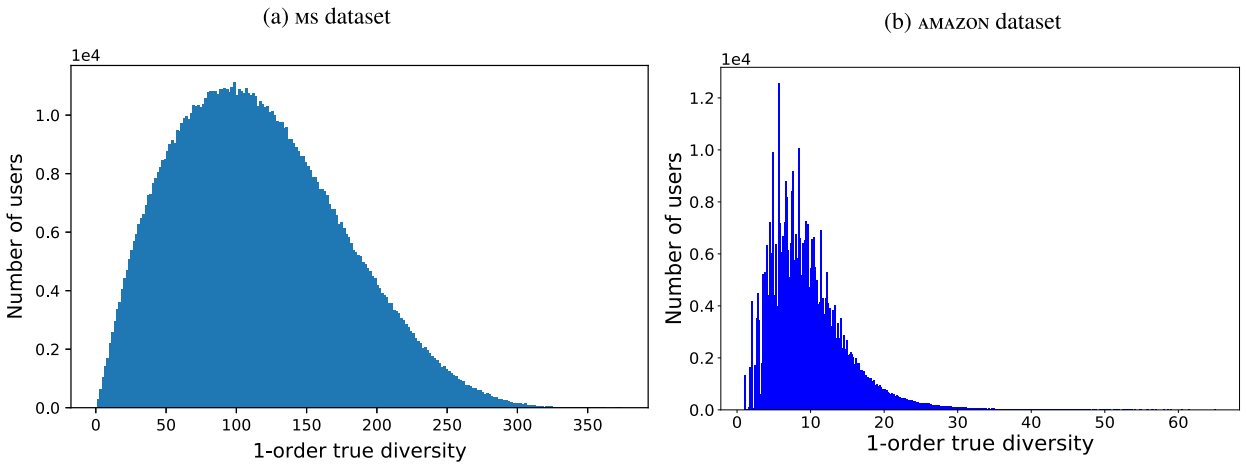


Fig. 11. Histograms showing the distributions of the 1-order true diversities of the attention of users ($D_1(p_{\Pi_{\text{attention}}|u})$) for users $u \in V_{\text{users}}$ for the MS (left) and AMAZON (right) datasets.

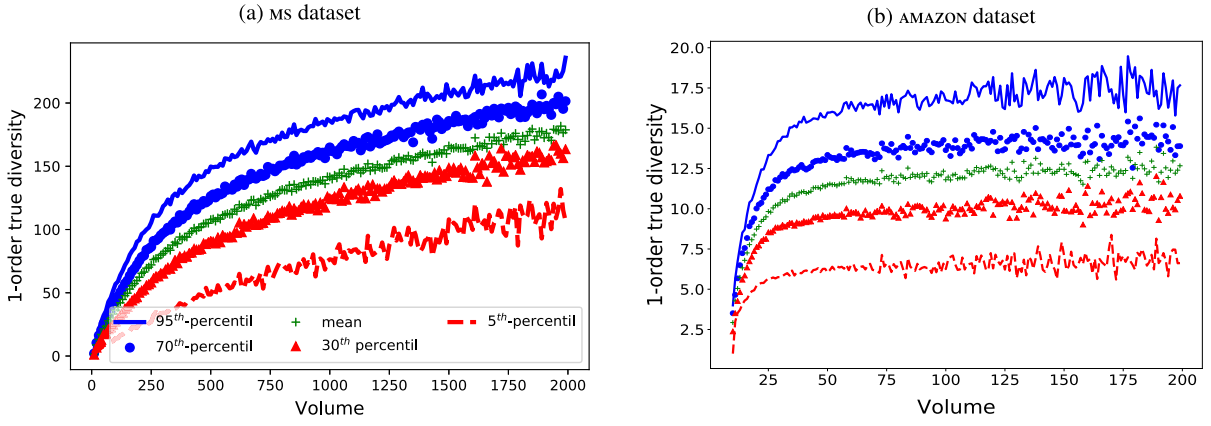


Fig. 12. Evolution of the 1-order true diversity of users' attention as a function of its volume for the *Million Song Dataset* (left) and *Amazon Dataset* (right).

However, it is possible to study how the diversity of the users depends on their activity on the platform. More precisely, let us define the *volume* of a user $u \in V_{\text{users}}$ as the sum of the number of tags for all songs listened to (in the case of the MS dataset) or reviewed (in the case of the AMAZON dataset) by u, multiplied by their play count (the number of songs consumed by user u). Then we can investigate whether there is a correlation between volume and diversity. Intuitively, the highest the volume, the highest the diversity: as its volume increases, a user has indeed more opportunities to explores new musical categories, thus diversifying its activity on the platform.

To see this more clearly, Fig. 12 presents the mean value of the 1-order true diversity of a user as a function of its volume, along with the 5th-, 30th-, 70th- and 95th-percentile. For both dataset, we can observe that the diversity increases along with the volume. However, we can also notice that the influence of the volume is clearly lower after a given threshold, highlighting a *saturation process* in the diversity of users' attention: while the growth of the diversity is initially sharp as the volume increases, after a given threshold (around 250 in MS and 25 in AMAZON), the users listen repeatedly to, or review similar contents proposed by the platform. This is particularly obvious in AMAZON (Fig. 12b) but one can also spot this phenomenon on MS (Fig. 12a).

After having presented a simple example of application of the network diversity measures in Section 5.1 and a numerical example with real datasets in this Section 5.2, we present, in the rest of Section 5 application examples for different research domains

5.3. Recommender systems

Diversity and diversification of algorithmic recommendations has become one of the leading topics of the recommender systems research community [114,115]. Through a variety of means, users have access today to large numbers 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 [116], 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 [117,118].

In recommender systems, diversity can help improve users' appreciation of the quality of recommendations [119,120]. It also has other applications, such as detecting changes in consumption behavior for context-aware recommenders [121]. 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 [122], or [119,120] for a discussion of terminology and definitions). These indicators are often computed using past collective choices of items made by users [115], 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 our 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. Fig. 13 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).

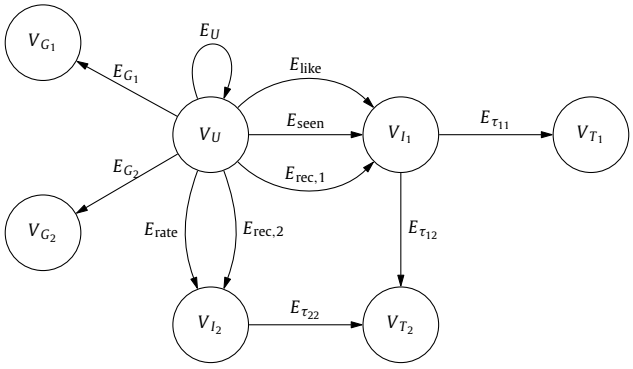


Fig. 13. Network schema of a heterogeneous information network in a setting from recommender systems, where users in V_U, belonging to groups V_{G_1} and V_{G_2} interact and are recommended two different sets of items V_{I_1} and V_{I_2}, which are classified using types in V_{T_1} and V_{T_2}.

In order to consider random walks constrained to meta 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 Fig. 13, we also consider different edge types:

- An edge type between users E_U (*e.g.*, users following or friending each other on a social network);
- Edge types from groups to users: E_{G_1} (*e.g.*, associating users with demographic groups) and E_{G_2} (*e.g.*, associating users with locations);
- Edge types indicating classification of items into types: $E_{\tau_{11}}$ and $E_{\tau_{12}}$ for V_{I_1}, and $E_{\tau_{22}}$ for V_{I_2};
- 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 ($E_{\text{rec},1}$ and $E_{\text{rec},2}$).

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 (V_{I_1} and V_{I_2} in our example) are common in cross-domain recommendation [123], and more generally in heterogeneous information network recommendation [26]. Relations between users and items can be of different kinds in recommendation settings: edges can be used to indicate that a user has rated an item in *explicit feedback* –or scoring, or noting– systems (E_{rate} in the example), or to indicate that a user has liked an item in *implicit feedback* systems (E_{like} in the example). Some recommender systems and diversity measures can

take into account whether a user has previously seen an item [124] (E_{seen} in the example). Also, settings where meta-data are associated with users are very common in demographic or location filtering [125], and are represented in the example by using vertex types V_{G_1} and V_{G_2}. Finally, edges between users signaling relations such as friendship of a user *following* another one on social networks (E_U in the example) may also be exploited for recommendations [126,127], and certainly in diversity computations. As stated before, Fig. 13 represents the network schema (cf. Definition 7) of our example.

Most diversity computations in recommender systems consist in 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 [20] (*e.g.*, genres for films). In the proposed framework, this concept would be captured by the individual diversity. Let us imagine that V_U are users, that V_{I_2} are films, and that V_{T_2} are film genres (*e.g.*, comedy, thriller, etc.). The individual genre (V_{T_2}) diversity of films (V_{I_2}) recommended to a user $u \in V_U$ is

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

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

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

which can be computed as a geometric mean by choosing $X_U \sim \text{Uniform}(V_U)$ for the starting 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 chosen or liked it (sometimes called *novelty* [128]). In the proposed framework, an aggregation of this quantity for items proposed to all users corresponds to the following network diversity measure:

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

More interestingly, other relevant quantities expressible as network diversity measures have no explicit expression in other frameworks of the literature. The clearest example is the comparison between the mean individual and collective recommended diversities: for example, $D_\alpha \left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \mid X_U \right)$ versus $D_\alpha \left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \right)$. Distinguishing between these two concepts (cf. Fig. 6) is important when taking interest in diversity beyond its use as a quality of recommendations; for example, when studying phenomena such as filter bubbles, which could manifest at some level of aggregation of users, while still having a high collective diversity. Some concepts at the core of Recommender Systems could also be expressed in our network diversity framework, such as the so-called *User-Based Collaborative Filtering* (see Section 4.2 of [122]):

$$D_\alpha \left(X_U \xrightarrow{E_{\text{rec},1}} V_{I_1} \xrightarrow{E_{\text{like}}^{\intercal}} X'_U \xrightarrow{E_{\text{like}}} V'_{I_1} \xrightarrow{E_{\tau_{11}}} V_{T_1} \right),$$

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. The reader is referred to [129] for an example of the application of the network diversity measures in conjunction with recommendation tasks. In the cited article, after presenting different experimental protocols and datasets known to the Recommender Systems community, the authors examine the performance of recommendations using the networks diversity measures, whose theoretical development and properties are the object of this article.

5.4. 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 [130,131]. The analysis of phenomena such as echo chambers and filter bubbles through the measurement of diversity of information consumption is an established practice [21,22,132,133,23]. The settings of different social media studies vary. Concrete examples are the study of the *Leave* and *Remain* Brexit campaigns on Twitter [134] and the exchange of information between US Democrats and Republicans on Facebook [135].

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_{I_1} according to types V_{T_1} relative to the corresponding collective diversity	$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \parallel X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \mid X_U\right)$
Collective diversity of recommended items V_{I_1} according types V_{T_1} relative to the distribution of types of liked times	$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1} \parallel X_U \xrightarrow{E_{\text{like}}} X_{I_1} \xrightarrow{E_{\tau_{11}}} X_{T_1}\right)$
Collective diversity of recommendations of items in V_{I_1} (*e.g.*, films) according to types V_{T_2} (*e.g.*, genres) relative to the one of V_{I_2} (*e.g.*, series)	$D_\alpha\left(X_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_2} \parallel X_U \xrightarrow{E_{\text{rec},2}} X_{I_2} \xrightarrow{E_{\tau_{22}}} X_{T_2}\right)$
Diversity of items V_{I_1} recommended to friends of $u \in V_U$ according to types V_{T_1}	$D_\alpha\left(X_U \xrightarrow{E_U} X'_U \xrightarrow{E_{\text{rec},1}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_1} \mid X_U = u\right)$
Diversity of items V_{I_1} liked by group $g \in V_{G_1}$ according to types V_{T_1}	$D_\alpha\left(X_{G_1} \xrightarrow{E^\intercal_{G_1}} X_U \xrightarrow{E_{\text{like}}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_1} \mid X_{G_1} = g\right)$
Diversity of users V_U that liked items V_{I_1} of type $t \in V_{T_2}$	$D_\alpha\left(X_U \mid X_U \xrightarrow{E_{\text{like}}} X_{I_1} \xrightarrow{E_{\tau_{12}}} X_{T_2} = t\right)$
Diversity of types V_{T_2} chosen by user $u \in V_U$ through their choices of items in V_{I_1}	$D_\alpha\left(X_U \xrightarrow{E\left(E_{\text{like}} \cdot E_{\tau_{12}}\right)} X_{T_2} \mid X_U = u\right)$

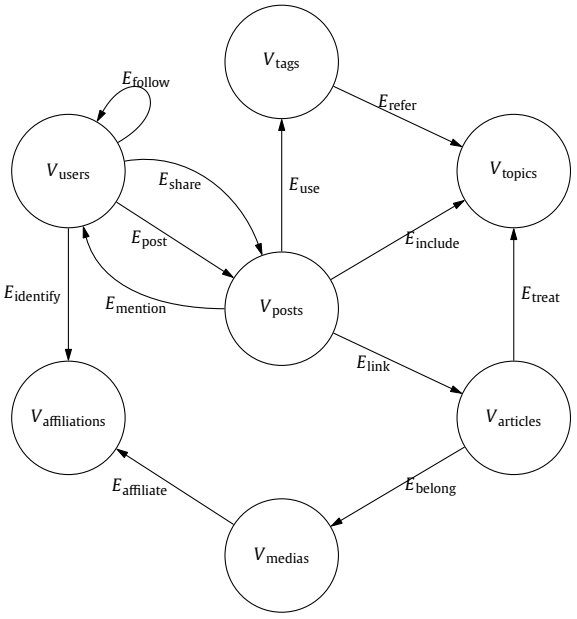


Fig. 14. Network schema of a heterogeneous information network in a setting from social networks and media studies, suited for the study of echo chambers and filter bubbles.

In this section, we illustrate the use of network diversity measures for the study of information exchange on social networks. We consider, as before, a heterogeneous information network created from activity traces of social networks and media. Fig. 14 shows its network schema, with which we may illustrate the use of network diversity measures in this context. In this example, we consider: users that post or share posts (or *tweets*, or blog entries, or comments in forums), users that can follow (or befriend) other users, posts that may 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 may be associated with media outlets, which may 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. [130] 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 [136], explicitly consider networks of information items (*e.g.*, blog posts) and the concepts that they use.

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	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}}\right)$
Collective affiliation diversity of users through the contents they share	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{share}}} X_{\text{posts}} \xrightarrow{E_{\text{link}}} X_{\text{articles}} \xrightarrow{E_{\text{belong}}} X_{\text{medias}} \xrightarrow{E_{\text{affiliate}}} X_{\text{affiliations}}\right)$
Individual topic diversity of user $u \in V_{\text{users}}$ through posting	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{include}}} X_{\text{topics}} \mid X_{\text{users}} = u\right)$
Individual topic diversity of user $u \in V_{\text{users}}$ through posting of followers	$D_\alpha\left(X_{\text{users}} \xrightarrow{E^{\intercal}_{\text{follow}}} X'_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{include}}} X_{\text{topics}} \mid X_{\text{users}} = u\right)$
Individual affiliation diversity of users mentioned by user $u \in V_{\text{users}}$	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{mention}}} X'_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}} \mid X_{\text{users}} = u\right)$
Diversity of affiliation groups that treat topic $t \in V_{\text{topics}}$ in articles	$D_\alpha\left(X_{\text{affiliations}} \mid X_{\text{affiliations}} \xrightarrow{E^{\intercal}_{\text{affiliate}}} X_{\text{medias}} \xrightarrow{E^{\intercal}_{\text{belong}}} X_{\text{articles}} \xrightarrow{E_{\text{treat}}} V_{\text{topics}} = t\right)$
Affiliation diversity of users according to who they mention in their posts relative to their own identified political affiliation	$D_\alpha\left(X_{\text{users}} \xrightarrow{E_{\text{post}}} X_{\text{posts}} \xrightarrow{E_{\text{mention}}} X'_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}} \parallel X_{\text{users}} \xrightarrow{E_{\text{identify}}} X_{\text{affiliations}}\right)$

As with the example of a generic recommender systems, we present in a schematic fashion (in Table 4) 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.

5.5. 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* [137]: 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 [138,139]. Similarly, other relations between species have been described using graphs, such as parasitation [140]. 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 [141,142].

We suggest that all these elements present in ecology can be treated using heterogeneous information networks. Using network diversity measures, different concepts related to diversity can be computed. Let us consider for example a heterogeneous information network with vertex types for habitats (V_{habitats}), for individuals ($V_{\text{individuals}}$), for species (V_{species}), for genera (V_{genera}), for families (V_{families}), and so on as needed.

Let us also consider for our example several edge types. Edge type E_{connect} is that of edges between habitats, indicating whether an individual can access a given habitat from another one. Edge type E_{inhabit} is used to represent which individuals inhabit which habitats. Edge types E_{eat} and E_{parasite} are used to represent relations between species; which species eat which species, and similarly for parasitation. Edge type $E_{\text{belong},1}$ is used to represent which individual belongs to which species. Finally, edge types $E_{\text{belong},2}$ and $E_{\text{belong},3}$ contain edges indicating which specie 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).

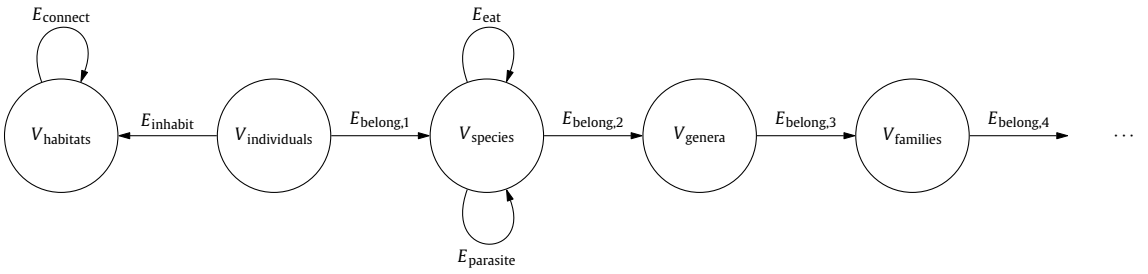
This setting can accommodate the common practices of measurement of –bio– diversity of a habitat $h \in V_{\text{habitats}}$, which in network diversity measures finds the expression

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

with $\alpha = 0$ giving the richness biodiversity, and $\alpha = \infty$ the Berger-Parker biodiversity of the habitat. Fig. 15 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.

5.6. 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



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

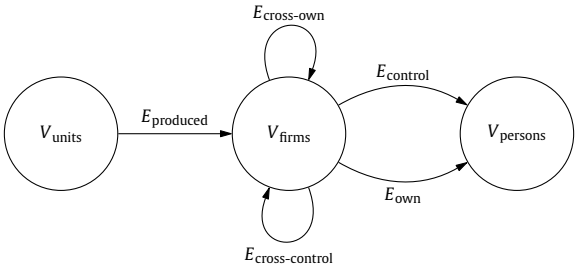
Fig. 15. Network schema of an example from ecology, and table with examples of diversity-related concepts and their expression as network diversity measures.

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 [67]. 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 [143,144], and that address the complex structure of co-ownership networks and their importance in regulation [145,146]. This makes graph-theoretical approaches good candidates for making advances in the measurement of concentration in industries [147]. 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 unit to the firm that produced them, E_{own}, linking firms to the persons that own them, $E_{\text{cross-own}}$, linking firms with each other according to cross-ownership, and similarly, E_{control} and $E_{\text{cross-control}}$ 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}, $E_{\text{cross-own}}$, E_{control}, and $E_{\text{cross-control}}$, the multiplicity of edges can account for the units by which property or control



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

Fig. 16. 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 other.

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 $E_{\text{cross-own}}$.

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

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

which becomes the Herfindahl-Hirschman Diversity (reciprocal of the Herfindahl-Hirschman Index) by choosing $\alpha = 2$. Fig. 16 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.

5.7. Scientometrics

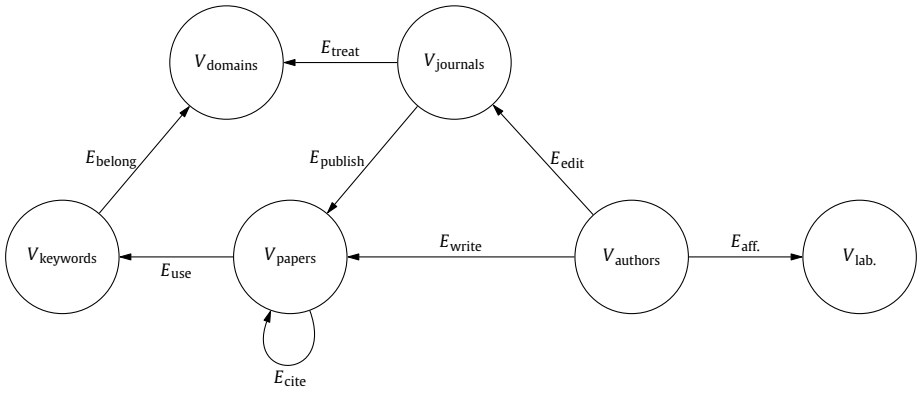
Scientometrics, within the field of bibliometrics, studies the measurement and analysis of scientific literature. Overlapping with information systems, scientometrics study, for example, the *importance* of publications in networks of citations using metrics such as the *Impact Factor* or the *Science 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 [148].

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 [149], and in general for the understanding and description of the structure of scientific communities [150,151]. Another practical application of the measurement of diversity in citation networks is the maintenance of classification systems [152].

Below, 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_{authors}, laboratories $V_{\text{lab.}}$ (or affiliation institutions), journals V_{journals}, scientific articles V_{papers}, keywords used by these articles V_{keywords}, and domains of research V_{domains} (*e.g.*, ecology, economics). We also consider edge types for representing relations between these entities (see Fig. 17): affiliation $E_{\text{aff.}}$ of authors to institutions, edition (or peer-review) E_{edit} of journals by authors, writing E_{write} of articles by authors, use of keywords E_{use} in articles, association of keywords E_{belong} with domains, publication E_{publish} of articles by journals, and declared treatment E_{treat} of research domains by journals. Fig. 17 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.



Examples of concepts expressible in research questions	Corresponding network diversity measures
Diversity of keywords used by author $a \in V_{\text{authors}}$	$D_\alpha \left(X_{\text{authors}} \xrightarrow{E_{\text{write}}} X_{\text{papers}} \xrightarrow{E_{\text{use}}} X_{\text{keywords}} \mid X_{\text{authors}} = a \right)$
Diversity of domains addressed in publications by author $a \in V_{\text{authors}}$ relative to domains he or she addresses in editing (or peer-reviewing)	$D_\alpha \left(X_{\text{authors}} \xrightarrow{E_{\text{write}}} X_{\text{papers}} \xrightarrow{E_{\text{use}}} X_{\text{keywords}} \xrightarrow{E_{\text{belong}}} X_{\text{domains}} \mid X_{\text{authors}} = a \parallel X'_{\text{authors}} \xrightarrow{E_{\text{edit}}} X_{\text{journals}} \xrightarrow{E_{\text{treat}}} X'_{\text{domains}} \mid X'_{\text{authors}} = a \right)$
Diversity of domains addressed by citations by authors of laboratory $l \in V_{\text{lab.}}$	$D_\alpha \left(X_{\text{lab.}} \xrightarrow{E_{\text{aff.}}^{\mathsf{T}}} X_{\text{authors}} \xrightarrow{E_{\text{write}}} X_{\text{papers}} \xrightarrow{E_{\text{cite}}} X'_{\text{papers}} \xrightarrow{E_{\text{use}}} X_{\text{keywords}} \xrightarrow{E_{\text{belong}}} X_{\text{domains}} \mid X_{\text{lab.}} = l \right)$

Fig. 17. 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.

By presenting a concise 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 α: the *true diversities*. Providing a formalism for heterogeneous information networks and constrained random walks on it, we consider different probability distributions on which diversity measures are computed. These diversity measures are related to the structure of the heterogeneous information network, and thus to the 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.
- 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.

In addition to providing means of referencing, expression, and computation on diversity-related concepts in complex data modeled by heterogeneous information networks, the *network diversity measures* could be leveraged in different downstream tasks performed in data mining. Expanding on those mentioned in the introduction (Section 1), we can now hint at more precise examples of applications in such tasks. In Recommender Systems, for example, previous works have used diversities associated with meta paths to avoid over-fitting in the training stage on data modeled with heterogeneous information networks [153]. Network diversity measures could allow, in this case, to target different collective and individual diversities in the function to be optimized. Other applications use meta paths to consider node similarities in strategies for node classification [154], and could leverage network diversity measures to consider parametrization of the weight given to

balance and variety in diversity when used. Research in representation learning could embed different quantities computable with network diversity measures in learning strategies where diversity would be leveraged for optimizing different objective function through training [155]. Finally, we hint to other applications that would need to be explored in greater detail than it is possible here, in domains related to heterogeneous information networks and diversity, such as community detection [156–158], clustering in networks [159,160], and the analysis of time-series resulting from temporal networks [161].

Many relevant advances in computer sciences hinge on the improvement of performance metrics through the development of novel algorithms and methodologies. This article seeks to contribute in the proposal of the metrics with which advancements are to be compared with the state of art. It is for this purpose what we propose a detailed discussion of the properties and implications of the proposed network diversity measures. 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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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, The choice between diversity and scale, in: E. Davis (Ed.), 1992: Myths and Realities, Centre for Business Strategy, London Business School, London, UK, 1989, pp. 29–45.
[3] J. Aczél, Z. Daróczy, On Measures of Information and Their Characterizations, 1975, p. 168, New York.
[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 Comput. Sci. 1 (2015) e38.
[6] J. Kulshrestha, M.B. Zafar, L.E. Noboa, K.P. Gummadi, S. Ghosh, Characterizing information diets of social media users, in: ICWSM, 2015, pp. 218–227.
[7] R.K. Ursem, Diversity-guided evolutionary algorithms, in: International Conference on Parallel Problem Solving from Nature, Springer, 2002, pp. 462–471.
[8] J. Riget, J.S. Vesterstrøm, A diversity-guided particle swarm optimizer - the ARPSO, Dept. Comput. Sci., Univ. of Aarhus, Aarhus, Denmark, Tech. Rep. 2 2002.
[9] A. Stirling, A general framework for analysing diversity in science, technology and society, J. R. Soc. Interface 4 (15) (2007) 707–719.
[10] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[11] C. Gini, Measurement of inequality of incomes, Econ. J. 31 (121) (1921) 124–126.
[12] S.A. Rhoades, The Herfindahl-Hirschman index, Fed. Reserve Bull. 79 (1993) 188.
[13] M.O. Hill, Diversity and evenness: a unifying notation and its consequences, Ecology 54 (2) (1973) 427–432.
[14] D. Encaoua, A. Jacquemin, Degree of monopoly, indices of concentration and threat of entry, Int. Econ. Rev. (1980) 87–105.
[15] S.R. Chakravarty, W. Eichhorn, An axiomatic characterization of a generalized index of concentration, J. Product. Anal. 2 (2) (1991) 103–112.
[16] F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information, Harvard University Press, Cambridge, MA, USA, 2015.
[17] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You, Penguin Books Limited, 2011.
[18] A. Datta, J. Makagon, D. Mulligan, M. Tschantz, Discrimination in online personalization: a multidisciplinary inquiry, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, ACM, 2018.
[19] R. Courtland, Bias detectives: the researchers striving to make algorithms fair, Nature 558 (7710) (2018) 357–360.
[20] C.-N. Ziegler, S.M. McNee, J.A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th International Conference on World Wide Web, ACM, 2005, pp. 22–32.
[21] B. Smyth, P. McClave, Similarity vs. diversity, in: Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, ICCBR '01, Springer-Verlag, 2001, pp. 347–361.
[22] A.-A. Stoica, C. Riederer, A. Chaintreau, Algorithmic glass ceiling in social networks: the effects of social recommendations on network diversity, in: Proceedings of the 2018 World Wide Web Conference, ACM, 2018, pp. 923–932.
[23] M. Schedl, P. Knees, F. Gouyon, New paths in music recommender systems research, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM, 2017, pp. 392–393.
[24] Y. Sun, J. Han, Mining heterogeneous information networks: a structural analysis approach, ACM SIGKDD Explor. Newsl. 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] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, J. Han, Personalized entity recommendation: a heterogeneous information network approach, in: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, ACM, 2014, pp. 283–292.
[27] 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.
[28] M. Sydow, M. Pikuła, R. Schenkel, The notion of diversity in graphical entity summarisation on semantic knowledge graphs, J. Intell. Inf. Syst. 41 (2) (2013) 109–149.

[29] Y. Chen, C. Wang, HINE: heterogeneous information network embedding, in: International Conference on Database Systems for Advanced Applications, Springer, 2017, pp. 180–195.
[30] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M.A. Porter, S. Gómez, A. Arenas, Mathematical formulation of multilayer networks, Phys. Rev. X 3 (4) (2013) 041022.
[31] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, A. Arenas, Ranking in interconnected multilayer networks reveals versatile nodes, Nat. Commun. 6 (2015) 6868.
[32] Y. Sun, J. Han, X. Yan, P.S. Yu, T. Wu, PathSim: meta path-based top-k similarity search in heterogeneous information networks, Proc. VLDB Endow. 4 (11) (2011) 992–1003.
[33] N. Lao, W.W. Cohen, Relational retrieval using a combination of path-constrained random walks, Mach. Learn. 81 (1) (2010) 53–67.
[34] Y. Xiong, Y. Zhu, S.Y. Philip, Top-K similarity join in heterogeneous information networks, IEEE Trans. Knowl. Data Eng. 27 (6) (2014) 1710–1723.
[35] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, J. Han, Recommendation in heterogeneous information networks with implicit user feedback, in: Proceedings of the 7th ACM Conference on Recommender Systems, ACM, 2013, pp. 347–350.
[36] C. Shi, B. Hu, W.X. Zhao, S.Y. Philip, Heterogeneous information network embedding for recommendation, IEEE Trans. Knowl. Data Eng. 31 (2) (2018) 357–370.
[37] K. Yu, S. Yu, V. Tresp, Soft clustering on graphs, in: Advances in Neural Information Processing Systems, 2006, pp. 1553–1560.
[38] R.-H. Li, J.X. Yu, Scalable diversified ranking on large graphs, IEEE Trans. Knowl. Data Eng. 25 (9) (2012) 2133–2146.
[39] D. Hristova, M.J. Williams, M. Musolesi, P. Panzarasa, C. Mascolo, Measuring urban social diversity using interconnected geo-social networks, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 21–30.
[40] B. King, R. Jha, D.R. Radev, Heterogeneous networks and their applications: scientometrics, name disambiguation, and topic modeling, Trans. Assoc. Comput. Linguist. 2 (2014) 1–14.
[41] S. Nandanwar, A. Moroney, M.N. Murty, Fusing diversity in recommendations in heterogeneous information networks, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM, 2018, pp. 414–422.
[42] F. Vahedian, R. Burke, B. Mobasher, Meta-path selection for extended multi-relational matrix factorization, in: The Twenty-Ninth International Flairs Conference, 2016.
[43] X. Yu, Y. Sun, B. Norick, T. Mao, J. Han, User guided entity similarity search using meta-path selection in heterogeneous information networks, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 2025–2029.
[44] G. Helfman, B.B. Collette, D.E. Facey, B.W. Bowen, The Diversity of Fishes: Biology, Evolution, and Ecology, John Wiley & Sons, 2009.
[45] S.J. McNaughton, Diversity and stability of ecological communities: a comment on the role of empiricism in ecology, Am. Nat. 111 (979) (1977) 515–525.
[46] R.M. May, Patterns of species abundance and diversity, Ecol. Evol. Commun. (1975) 81–120.
[47] J.M. Smith, Trees, bundles or nets? Trends Ecol. Evol. 4 (10) (1989) 302–304.
[48] S. Gillett, In praise of policy diversity, Position paper for OII broadband forum, 2003.
[49] G. Silverberg, G. Dosi, L. Orsenigo, Innovation, diversity and diffusion: a self-organisation model, Econ. J. 98 (393) (1988) 1032–1054.
[50] Helga Nowotny, Peter B. Scott, Michael T. Gibbons, Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty, John Wiley & Sons, 2013.
[51] E.V. Shevchenko, D.V. Talapin, N.A. Kotov, S. O'Brien, C.B. Murray, Structural diversity in binary nanoparticle superlattices, Nature 439 (7072) (2006) 55.
[52] E.D. Schneider, J.J. Kay, Life as a manifestation of the second law of thermodynamics, Math. Comput. Model. 19 (6–8) (1994) 25–48.
[53] G. Grabher, D. Stark, Organizing diversity: evolutionary theory, network analysis and postsocialism, Reg. Stud. 31 (5) (1997) 533–544.
[54] W.R. Ashby, Requisite variety and its implications for the control of complex systems, in: Facets of Systems Science, Springer, 1991, pp. 405–417.
[55] F.J. Dyson, Statistical theory of the energy levels of complex systems. I, J. Math. Phys. 3 (1) (1962) 140–156.
[56] H.-X. Yang, Z.-X. Wu, C. Zhou, T. Zhou, B.-H. Wang, Effects of social diversity on the emergence of global consensus in opinion dynamics, Phys. Rev. E 80 (4) (2009) 046108.
[57] A. Stirling, On the economics and analysis of diversity, Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper 28, 1998, pp. 1–156.
[58] R.H. MacArthur, Patterns of species diversity, Biol. Rev. 40 (4) (1965) 510–533.
[59] N.J. Gotelli, R.K. Colwell, Estimating species richness, in: Biological Diversity: Frontiers in Measurement and Assessment, vol. 12, 2011, pp. 39–54.
[60] E.P. Odum, Fundamentals of Ecology, WB Saunders Company, 1959.
[61] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL, 1963.
[62] E.C. Pielou, et al., An Introduction to Mathematical Ecology, Wiley, New York, NY, 1969.
[63] M. Wulfmeier, P. Ondruska, I. Posner, Maximum entropy deep inverse reinforcement learning, arXiv preprint, arXiv:1507.04888, 2015.
[64] W. Wang, G. Zhang, J. Lu, Collaborative filtering with entropy-driven user similarity in recommender systems, Int. J. Intell. Syst. 30 (8) (2015) 854–870.
[65] E.H. Simpson, Measurement of diversity, Nature (1949).
[66] A.O. Hirschman, National Power and the Structure of Foreign Trade, University of California Press, 1945.
[67] O.C. Herfindahl, Concentration in the US steel industry, Unpublished PhD. Dissertation, Columbia University, 1950.
[68] J.P. Gibbs, W.T. Martin, Urbanization, technology, and the division of labor: international patterns, Am. Sociol. Rev. (1962) 667–677.
[69] M. Nei, Estimation of average heterozygosity and genetic distance from a small number of individuals, Genetics 89 (3) (1978) 583–590.
[70] S.H. Hurlbert, The nonconcept of species diversity: a critique and alternative parameters, Ecology 52 (4) (1971) 577–586.
[71] A. Sen, M.A. Sen, S. Amartya, J.E. Foster, J.E. Foster, et al., On Economic Inequality, Oxford University Press, 1997.
[72] W.H. Berger, F.L. Parker, Diversity of planktonic foraminifera in deep-sea sediments, Science 168 (3937) (1970) 1345–1347.
[73] M.L. Weitzman, On diversity, Q. J. Econ. 107 (2) (1992) 363–405.
[74] A.R. Solow, S. Polasky, Measuring biological diversity, Environ. Ecol. Stat. 1 (2) (1994) 95–103.
[75] P.H. Williams, C. Humphries, R. Vane-Wright, Measuring biodiversity: taxonomic relatedness for conservation priorities, Aust. Syst. Bot. 4 (4) (1991) 665–679.
[76] P. Nguyen, P.-P. Saviotti, M. Trommetter, B. Bourgeois, Variety and the evolution of refinery processing, Ind. Corp. Change 14 (3) (2005) 469–500.
[77] B. Runnegar, K. Campbell, M. Day, Rates and Modes of Evolution in the Mollusca, Rates of Evolution, Allen and Unwin, London, 1987, pp. 39–60.
[78] K. Junge, Diversity of ideas about diversity measurement, Scand. J. Psychol. 35 (1) (1994) 16–26.
[79] C.R. Rao, Rao's axiomatization of diversity measures, in: Wiley StatsRef: Statistics Reference Online, 2014.
[80] I. Csiszár, Axiomatic characterizations of information measures, Entropy 10 (3) (2008) 261–273.
[81] G. Wang, M. Jiang, Axiomatic characterization of nonlinear homomorphic means, J. Math. Anal. Appl. 303 (1) (2005) 350–363.
[82] J. Aczél, C. Alsina, Synthesizing judgements: a functional equations approach, Math. Model. 9 (3–5) (1987) 311–320.
[83] H. Dalton, The measurement of the inequality of incomes, Econ. J. 30 (119) (1920) 348–361.
[84] S. Hoffmann, Concavity and additivity in diversity measurement: re-discovery of an unknown concept, Univ., FEMM, 2006.
[85] A.N. Kolmogorov, G. Castelnuovo, Sur la notion de la moyenne, G. Bardi, tip. della R. Accad. dei Lincei, 1930.
[86] M. Nagumo, Über eine Klasse der Mittelwerte, Jpn. J. Math., Trans. Abstr. 7 (1930) 71–79, The Mathematical Society of Japan.

[87] H. Pursiainen, Consistency in aggregation, quasilinear means and index numbers, Quasilinear Means and Index Numbers, 2008.
[88] M. Hall, N. Tideman, Measures of concentration, J. Am. Stat. Assoc. 62 (317) (1967) 162–168.
[89] L. Jost, Entropy and diversity, Oikos 113 (2) (2006) 363–375.
[90] L. Hannah, J.A. Kay, Concentration in Modern Industry: Theory, Measurement and the UK Experience, Springer, 1977.
[91] L. Jost, P. DeVries, T. Walla, H. Greeney, A. Chao, C. Ricotta, Partitioning diversity for conservation analyses, Divers. Distrib. 16 (1) (2010) 65–76.
[92] A.J. Daly, J.M. Baetens, B. De Baets, Ecological diversity: measuring the unmeasurable, Mathematics 6 (7) (2018) 119.
[93] P. Ramaciotti Morales, L. Tabourier, S. Ung, C. Prieur, Role of the website structure in the diversity of browsing behaviors, in: Proceedings of the 30th ACM Conference on Hypertext and Social Media, ACM, 2019, pp. 133–142.
[94] T. Van Erven, P. Harremos, Rényi divergence and Kullback-Leibler divergence, IEEE Trans. Inf. Theory 60 (7) (2014) 3797–3820.
[95] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[96] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley Series in Telecommunications and Signal Processing, Wiley-Interscience, New York, NY, USA, 2006.
[97] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, 2008, pp. 1247–1250.
[98] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[99] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25.
[100] C. Shi, Y. Li, J. Zhang, Y. Sun, S.Y. Philip, A survey of heterogeneous information network analysis, IEEE Trans. Knowl. Data Eng. 29 (1) (2016) 17–37.
[101] A. Singhal, Introducing the knowledge graph: things, not strings, Official Google blog 5, 2012.
[102] W.W.W. Consortium, Resource description framework (RDF), https://www.w3.org/RDF/.
[103] J. Han, J.-R. Wen, Mining frequent neighborhood patterns in a large labeled graph, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013, pp. 259–268.
[104] M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, A review of relational machine learning for knowledge graphs, Proc. IEEE 104 (1) (2015) 11–33.
[105] Y. Zheng, C. Shi, X. Cao, X. Li, B. Wu, Entity set expansion with meta path in knowledge graph, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2017, pp. 317–329.
[106] X. Cao, C. Shi, Y. Zheng, J. Ding, X. Li, B. Wu, A heterogeneous information network method for entity set expansion in knowledge graph, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 288–299.
[107] N. Lao, W.W. Cohen, Fast query execution for retrieval models based on path-constrained random walks, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 881–888.
[108] C. Shi, S.Y. Philip, Heterogeneous Information Network Analysis and Applications, Springer, 2017.
[109] R. Poulain, F. Tarissan, Investigating the lack of diversity in user behavior: the case of musical content on online platforms, Inf. Process. Manag. 57 (2) (2020) 102169.
[110] T. Bertin-Mahieux, D.P. Ellis, B. Whitman, P. Lamere, The million song dataset, in: Proceedings of the 12th International Conference on Music Information Retrieval, University of Miami, 2011, pp. 591–596.
[111] R. He, J. McAuley, Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering, in: Proceedings of the 25th International Conference on World Wide Web, ACM, 2016, pp. 507–517.
[112] J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2015, pp. 43–52.
[113] I. Peters, Folksonomies. Indexing and Retrieval in Web 2.0, Walter de Gruyter, 2009.
[114] M. Kunaver, T. Požrl, Diversity in recommender systems–a survey, Knowl.-Based Syst. 123 (2017) 154–162.
[115] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J.R. Wakeling, Y.-C. Zhang, Solving the apparent diversity-accuracy dilemma of recommender systems, Proc. Natl. Acad. Sci. 107 (10) (2010) 4511–4515.
[116] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
[117] S.M. McNee, J. Riedl, J.A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in: CHI'06 Extended Abstracts on Human Factors in Computing Systems, ACM, 2006, pp. 1097–1101.
[118] K. Bradley, B. Smyth, Improving recommendation diversity, in: Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland, Citeseer, 2001, pp. 85–94.
[119] T. Silveira, M. Zhang, X. Lin, Y. Liu, S. Ma, How good your recommender system is? A survey on evaluations in recommendation, Int. J. Mach. Learn. Cybern. 10 (5) (2019) 813–831.
[120] J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez, Recommender systems survey, Knowl.-Based Syst. 46 (2013) 109–132.
[121] A. L'Huillier, S. Castagnos, A. Boyer, Modéliser la diversité au cours du temps pour détecter le contexte dans un service de musique en ligne, in: Revue des Sciences et Technologies del'Information - Série TSI: Technique et Science Informatiques, Lavoisier, 2016, hal-01300419.
[122] P. Kantor, F. Ricci, L. Rokach, B. Shapira, Recommender Systems Handbook, Springer, 2010.
[123] J. Tang, S. Wu, J. Sun, H. Su, Cross-domain collaboration recommendation, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1285–1293.
[124] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: Proceedings of the Fifth ACM Conference on Recommender Systems, ACM, 2011, pp. 109–116.
[125] L. Safoury, A. Salah, Exploiting user demographic attributes for solving cold-start problem in recommender system, Lect. Notes Softw. Eng. 1 (3) (2013) 303–307.
[126] G. Groh, C. Ehmig, Recommendations in taste related domains: collaborative filtering vs. social filtering, in: Proceedings of the 2007 International ACM Conference on Supporting Group Work, ACM, 2007, pp. 127–136.
[127] D. Bernardes, M. Diaby, R. Fournier, F. FogelmanSoulié, E. Viennet, A social formalism and survey for recommender systems, ACM SIGKDD Explor. Newsl. 16 (2) (2015) 20–37.
[128] N. Hurley, M. Zhang, Novelty and diversity in top-n recommendation–analysis and evaluation, ACM Trans. Internet Technol. 10 (4) (2011) 14.
[129] P. Ramaciotti Morales, L. Tabourier, R. Fournier-S'Niehotta, Testing the impact of semantics and structure on recommendation accuracy and diversity, in: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM, 2020.
[130] N. Gaumont, M. Panahi, D. Chavalarias, Reconstruction of the socio-semantic dynamics of political activist Twitter networks—method and application to the 2017 French presidential election, PLoS ONE 13 (9) (2018) e0201879.
[131] S. Flaxman, S. Goel, J.M. Rao, Filter bubbles, echo chambers, and online news consumption, Public Opin. Q. 80 (S1) (2016) 298–320.
[132] C. Sha, X. Wu, J. Niu, A framework for recommending relevant and diverse items, in: IJCAI, 2016, pp. 3868–3874.
[133] M.D. Ekstrand, M. Tian, M.R.I. Kazi, H. Mehrpouyan, D. Kluver, Exploring author gender in book rating and recommendation, in: Proceedings of the 12th ACM Conference on Recommender Systems, 2018, pp. 242–250.
[134] M. Bastos, D. Mercea, A. Baronchelli, The geographic embedding of online echo chambers: evidence from the Brexit campaign, PLoS ONE 13 (11) (2018) e0206841.

[135] E. Bakshy, S. Messing, L.A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, Science 348 (6239) (2015) 1130–1132.
[136] C. Roth, Socio-semantic frameworks, Adv. Complex Syst. 16 (04n05) (2013) 1350013.
[137] R.T. Paine, Food web complexity and species diversity, Am. Nat. 100 (910) (1966) 65–75.
[138] R.M. May, Ecology: the structure of food webs, Nature 301 (5901) (1983) 566–568.
[139] 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.
[140] R. Poulin, Network analysis shining light on parasite ecology and diversity, Trends Parasitol. 26 (10) (2010) 492–498.
[141] D. Urban, T. Keitt, Landscape connectivity: a graph-theoretic perspective, Ecology 82 (5) (2001) 1205–1218.
[142] M. Wirth, G.F. Estabrook, D.J. Rogers, A graph theory model for systematic biology, with an example for the Oncidiinae (Orchidaceae), Syst. Zool. 15 (1) (1966) 59–69.
[143] C. Kang, Ownership structure of Samsung chaebol, in: T. Shiba, M. Shimotani (Eds.), Beyond the Firm: Business Groups in International and Historical Perspective, 1997, pp. 31–58.
[144] H.-J. Kim, Concentrated ownership and corporate control: Wallenberg sphere and Samsung group, J. Korean Law 14 (2014) 39.
[145] H. Compston, The network of global corporate control: implications for public policy, Bus. Polit. 15 (3) (2013) 357–379.
[146] S. Vitali, J.B. Glattfelder, S. Battiston, The network of global corporate control, PLoS ONE 6 (10) (2011) e25995.
[147] M. Levy, Control in pyramidal structures, Corp. Gov. 17 (1) (2009) 77–89.
[148] E. Yan, Y. Ding, C.R. Sugimoto, P-rank: an indicator measuring prestige in heterogeneous scholarly networks, J. Am. Soc. Inf. Sci. Technol. 62 (3) (2011) 467–477.
[149] M. Zitt, E. Bassecoulard, Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues, Ethics Sci. Environ. Polit. 8 (1) (2008) 49–60.
[150] M. Zitt, Facing diversity of science: a challenge for bibliometric indicators, Measurement 3 (1) (2005) 38–49.
[151] G. Heimeriks, M. Hörlesberger, P. Van den Besselaar, Mapping communication and collaboration in heterogeneous research networks, Scientometrics 58 (2) (2003) 391–413.
[152] I. Gómez, M. Bordons, M. Fernandez, A. Méndez, Coping with the problem of subject classification diversity, Scientometrics 35 (2) (1996) 223–235.
[153] H. Liu, Z. Jiang, Y. Song, T. Zhang, Z. Wu, User preference modeling based on meta paths and diversity regularization in heterogeneous information networks, Knowl.-Based Syst. 181 (2019) 104784.
[154] D. Yin, H. Gao, A flexible aggregation framework on large-scale heterogeneous information networks, J. Inf. Sci. 43 (2) (2017) 186–203.
[155] C. Zhang, A. Swami, N.V. Chawla Shne, Representation learning for semantic-associated heterogeneous networks, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 690–698.
[156] J.D. Cruz, C. Bothorel, F. Poulet, Entropy based community detection in augmented social networks, in: 2011 International Conference on Computational Aspects of Social Networks (CASoN), IEEE, 2011, pp. 163–168.
[157] A.G. Nikolaev, R. Razib, A. Kucheriya, On efficient use of entropy centrality for social network analysis and community detection, Soc. Netw. 40 (2015) 154–162.
[158] K.R. Žalik, B. Žalik, Memetic algorithm using node entropy and partition entropy for community detection in networks, Inf. Sci. 445 (2018) 38–49.
[159] Y. Wang, F.S. Bao, An entropy-based weighted clustering algorithm and its optimization for Ad Hoc networks, in: Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2007), White Plains, NY, 2007, p. 56.
[160] E.C. Kenley, Y.-R. Cho, Entropy-based graph clustering: application to biological and social networks, in: 2011 IEEE 11th International Conference on Data Mining, IEEE, 2011, pp. 1116–1121.
[161] Y. Gu, A. McCallum, D. Towsley, Detecting anomalies in network traffic using maximum entropy estimation, in: Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, 2005, p. 32.