Axiomatic Characterisation of Diversity Measures on Tripartite Graphs

Robin Lamarche-Perrin[1,2], Lionel Tabourier[2], Fabien Tarissan[2,3], Raphaël Fournier-S'niehotta[4], and Rémy Cazabet[2]

[1] Institut des systèmes complexes de Paris Île-de-France
[2] Laboratoire d'informatique de Paris 6
[3] Institut des sciences sociales du politique
[4] Conservatoire national des arts et métiers

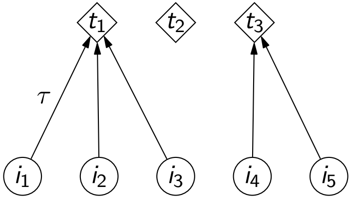
Second meeting of the ANR AlgoDiv Project
28th of June, 2017, in Paris

Starting point: Wikipedia article on "Diversity index"
https://en.wikipedia.org/wiki/Diversity_index

- A set of items $\mathcal{I} = \{i_1, \ldots, i_n\}$
- A set of type $\mathcal{T} = \{t_1, \ldots, t_m\}$
- A *membership function* $\tau : \mathcal{I} \to \mathcal{T}$

  defining the *proportional abundance* of types:

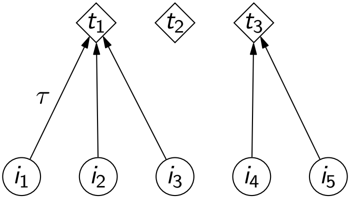
$$\forall t \in \mathcal{T}, \quad p_\tau(t) = \frac{|i \in \mathcal{I} : \tau(i) = t|}{|\mathcal{I}|}$$



$p_\tau(t_1) = \frac{3}{5}, \ p_\tau(t_2) = 0, \ p_\tau(t_3) = \frac{2}{5}$

A *diversity index* is an application $D : (\mathcal{I} \to \mathcal{T}) \to \mathbb{R}$ that associate a diversity value $D(\tau)$ to any membership function τ.

Disclaimer: Preliminary definitions needed!



The proper definition of any diversity measure hence requires to first solve three definition-related problems:

- **Identification problem:** Given a set of items $\mathcal{I}$...
- **Categorisation problem:** Given a set of types $\mathcal{T}$...
- **Enumeration problem:** How do we build the ratios $p_\tau(t)$?

In the following, we will assume these problems to be solved, and will focus on proper diversity measures that can be built on it.

Diversity measures 101

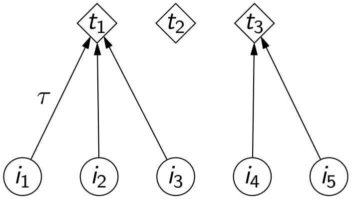
Starting point: Wikipedia article on “Diversity index”
https://en.wikipedia.org/wiki/Diversity_index

- **Richness:**

$$|t \in \mathcal{T} : p_\tau(t) > 0| = 2$$

- **Berger-Parker index:**

$$\max_{t \in \mathcal{T}} p_\tau(t) = \frac{3}{5}$$



$$p_\tau(t_1) = \tfrac{3}{5},\ p_\tau(t_2) = 0,\ p_\tau(t_3) = \tfrac{2}{5}$$

- **Herfindahl index:**

$$\sum_{t \in \mathcal{T}} p_\tau(t)^2 = 0.52$$

- **Shannon entropy:**

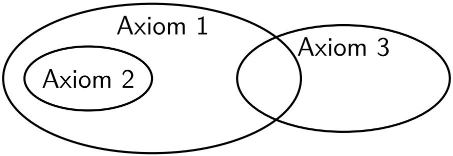
$$-\sum_{t \in \mathcal{T}} p_\tau(t) \log_2 p_\tau(t) \approx 0.971$$

Axiomatisation (1/2)

Coming from measures of diversity, of concentration, of inequality, of uncertainty...

- **Homogeneity:** $D(\tau) = D(p_\tau(t_1), \dots, p_\tau(t_m))$
- **Continuity:** $D(p_\tau(t_1), \dots, p_\tau(t_m))$ is a continuous function of $p_\tau(t_i)$
- **Symmetry:** $\forall \sigma \in \text{Sym}\{1, \dots, m\}, \quad D(p_1, \dots, p_m) = D(p_{\sigma(1)}, \dots, p_{\sigma(m)})$
- **Expansibility:** $D(p_1, \dots, p_m) = D(p_1, \dots, p_m, 0)$
- **Merging:** $D(\dots, p_i, \dots, p_j, \dots) \geq D(\dots, p_i + p_j, \dots)$
- **Monotonicity:** $\forall m_1 < m_2, \quad D\left(\frac{1}{m_1}, \dots, \frac{1}{m_1}\right) \leq D\left(\frac{1}{m_2}, \dots, \frac{1}{m_2}\right)$
- **Minimum:** $D(p_1, \dots, p_m) \geq D(0, \dots, 0, 1, 0, \dots, 0)$
- **Maximum:** $D(p_1, \dots, p_m) \leq D\left(\frac{1}{m}, \dots, \frac{1}{m}\right)$
- **Normalisation:** $D\left(\frac{1}{m}, \dots, \frac{1}{m}\right) = m$

Axiomatisation (2/2)



Assuming **homogeneity** and **symmetry**:

- **transfer principle** $\Rightarrow$ **maximum** and **Lorenz criterion**
- **maximum** and **merging** $\Rightarrow$ **monotonicity**
- **expansibility** and **transfer principle** $\Rightarrow$ **monotonicity**
- **normalisation** and **transfer principle** $\Rightarrow$ **minimum** and **maximum**

Encaoua and Jacquemin. 1980. "Degree of Monopoly, Indices of Concentration and Threat of Entry". In *International Economic Review*, vol. 21, n°1, p. 87-105.

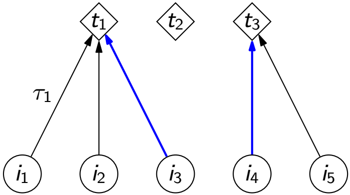
Chakravarty and Eichhorn. 1991. "An Axiomatic Characterization of a Generalized Index of Concentration". In *Journal of Productivity Analysis*, vol. 2, p. 103-112.

Homogeneity Axiom

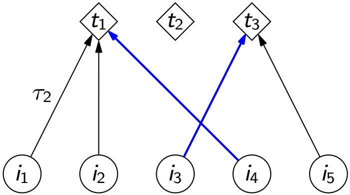
Homogeneity: Diversity measures should be neutral with respect to items, and hence should only depend on their distribution among types.

$$D(\tau) = D(p_\tau(t_1), \ldots, p_\tau(t_m))$$

E.g., swapping the types of two items should not change the system's diversity:



$p_{\tau_1}(t_1) = \frac{3}{5},\ p_{\tau_1}(t_2) = 0,\ p_{\tau_1}(t_3) = \frac{2}{5}$



$p_{\tau_2}(t_1) = \frac{3}{5},\ p_{\tau_2}(t_2) = 0,\ p_{\tau_2}(t_3) = \frac{2}{5}$

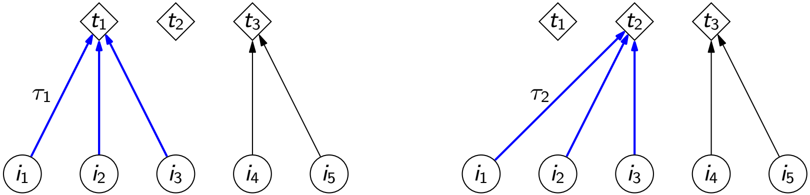
$\longrightarrow$ **Same diversity!**

Symmetry Axiom

Symmetry: Diversity measures should be neutral with respect to types, and hence should not depend on the ordering of types.

$$\forall \sigma \in \mathrm{Sym}\{1, \ldots, m\}, \quad D(p_1, \ldots, p_m) = D(p_{\sigma(1)}, \ldots, p_{\sigma(m)})$$

E.g., swapping two types should not change the system's diversity:



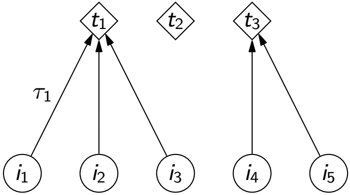
$p_{\tau_1}(t_1) = \frac{3}{5},\ p_{\tau_1}(t_2) = 0,\ p_{\tau_1}(t_3) = \frac{2}{5}$ $\quad$ $p_{\tau_2}(t_1) = 0,\ p_{\tau_2}(t_2) = \frac{3}{5},\ p_{\tau_2}(t_3) = \frac{2}{5}$

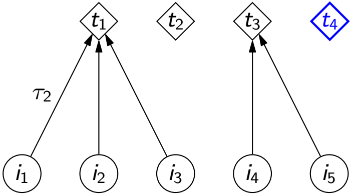
$\longrightarrow$ **Same diversity!**

Expansibility Axiom

Expansibility: Adding an empty type should not change the system's diversity.

$$D(p_1, \dots, p_m) = D(p_1, \dots, p_m, 0)$$





$p_{\tau_1}(t_1) = \frac{3}{5},\ p_{\tau_1}(t_2) = 0,\ p_{\tau_1}(t_3) = \frac{2}{5}$

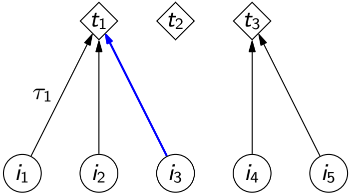
$p_{\tau_2}(t_1) = \frac{3}{5},\ p_{\tau_2}(t_2) = 0,\ p_{\tau_2}(t_3) = \frac{2}{5},\ p_{\tau_2}(t_4) = 0$

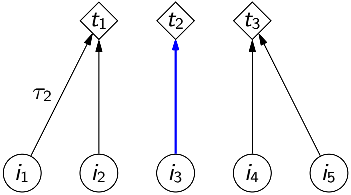
$\longrightarrow$ **Same diversity!**

Transfer Principle Axiom

Transfer principle: Increasing a rare type while decreasing a frequent type should increase the system's diversity.

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





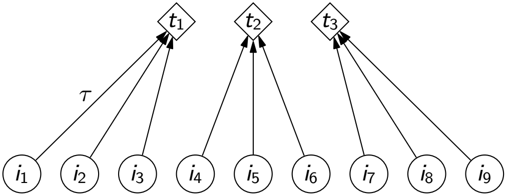
$p_{\tau_1}(t_1) = \frac{3}{5},\ p_{\tau_1}(t_2) = 0,\ p_{\tau_1}(t_3) = \frac{2}{5}$ $\qquad$ $p_{\tau_2}(t_1) = \frac{2}{5},\ p_{\tau_2}(t_2) = \frac{1}{5},\ p_{\tau_2}(t_3) = \frac{2}{5}$

$\longrightarrow$ **Diversity increase**

Normalisation Axiom

Normalisation: The diversity of a system containing equally-populated types should be given by the number of types.

$$D\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) = m$$



$$p_\tau(t_1) = \frac{1}{3},\ p_\tau(t_2) = \frac{1}{3},\ p_\tau(t_3) = \frac{1}{3}$$

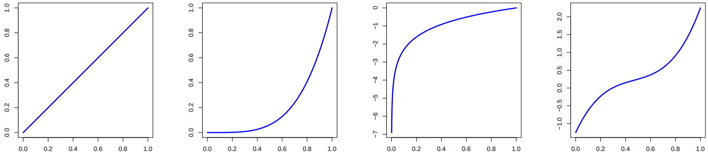
$\longrightarrow$ **Diversity = 3**

Assuming **homogeneity** and **symmetry**, which measures simultaneously satisfy the three following axioms: **expansibility**, **normalization**, and **transfert principle**?

Class of self-weighted quasilinear means:

$$D_\phi(\tau) = \phi^{-1}\left(\sum_{t \in \mathcal{T}} p_\tau(t)\ \phi(p_\tau(t))\right)$$

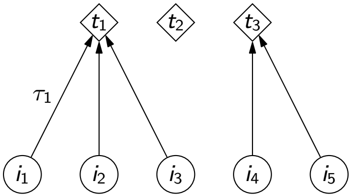
with $\phi : [0,1] \to \mathbb{R}$ continuous and strictly monotonic

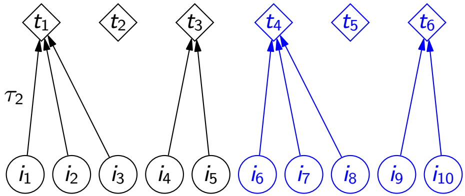


Replication Principle Axiom

Replication principle: Replicating the system k times should multiply its diversity by k.

$$D\Big(\underbrace{\frac{p_1}{k},\ldots,\frac{p_m}{k}}_{1^{st}},\ldots,\underbrace{\frac{p_1}{k},\ldots,\frac{p_m}{k}}_{k^{th}}\Big) = k\ D(p_1,\ldots,p_m)$$





$p_{\tau_1}(t_1) = \frac{3}{5},\ p_{\tau_1}(t_2) = 0,\ p_{\tau_1}(t_3) = \frac{2}{5}$

$p_{\tau_2}(t_1) = \frac{3}{10},\ p_{\tau_2}(t_2) = 0,\ p_{\tau_2}(t_3) = \frac{2}{10},$
$p_{\tau_2}(t_4) = \frac{3}{10},\ p_{\tau_2}(t_5) = 0,\ p_{\tau_2}(t_6) = \frac{2}{10}$

$\longrightarrow$ **Diversity** $\times 2$

More constrained class of measures

By adding the **replication principle**, then $\phi(x) = a\,x^{\alpha-1} + b$ with $\alpha > 0$

True diversity:

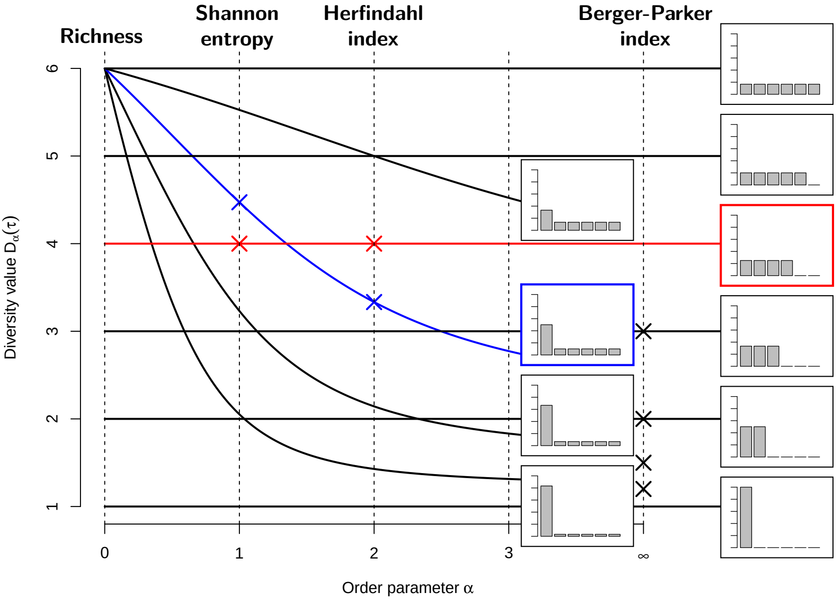
$$D_\alpha(\tau) = \left(\sum_{t \in \mathcal{T}} p_\tau(t)^\alpha \right)^{\frac{1}{1-\alpha}} \quad \text{with order parameter } \alpha \geq 0$$

$\alpha = 0 \quad \Rightarrow \quad D_0(\tau) = |t \in \mathcal{T} : p_\tau(t) > 0|$ **Richness**

$\alpha \to 1 \quad \Rightarrow \quad D_1(\tau) = \left(\prod_{t \in \mathcal{T}} p_\tau(t)^{p_\tau(t)} \right)^{-1}$ Exponential of **Shannon entropy**

$\alpha = 2 \quad \Rightarrow \quad D_2(\tau) = \left(\sum_{t \in \mathcal{T}} p_\tau(t)^2 \right)^{-1}$ Inverse of **Herfindahl index**

$\alpha \to \infty \quad \Rightarrow \quad D_\infty(\tau) = \left(\max_{t \in \mathcal{T}} p_\tau(t) \right)^{-1}$ Inverse of **Berger-Parker index**


Richness
Shannon entropy
Herfindahl index
Berger-Parker index
Diversity value $D_\alpha(\tau)$
Order parameter α
0
1
2
3
∞
1
2
3
4
5
6

Why Shannon entropy is special (1/2)

All true diversity measures verify the **weak additivity** axiom.

Weak additivity: If two typologies are *independent*, the diversity of their product is equal to the product of their diversity.

$$p_{(\tau_1,\tau_2)}(t_1, t_2) = p_{\tau_1}(t_1)\ p_{\tau_2}(t_2) \quad \Rightarrow \quad D(\tau_1, \tau_2) = D(\tau_1)\ D(\tau_2)$$

(τ_1, τ_2)	Comedy	Action	Drama	⋮	⋮		τ_2
Novel	$\frac{2}{40}$	$\frac{4}{40}$	0	$\frac{2}{40}$	$\frac{2}{40}$		$\frac{10}{40}$
Movie	$\frac{1}{40}$	$\frac{2}{40}$	0	$\frac{1}{40}$	$\frac{1}{40}$		$\frac{5}{40}$
Comics	$\frac{3}{40}$	$\frac{6}{40}$	0	$\frac{3}{40}$	$\frac{3}{40}$		$\frac{15}{40}$
…	0	0	0	0	0		0
…	$\frac{2}{40}$	$\frac{4}{40}$	0	$\frac{2}{40}$	$\frac{2}{40}$		$\frac{10}{40}$
τ_1	$\frac{8}{40}$	$\frac{16}{40}$	0	$\frac{8}{40}$	$\frac{8}{40}$		

Why Shannon entropy is special (2/2)

> Only Shannon entropy ($\alpha = 1$) verifies the **strong additivity** axiom.

Strong additivity: The diversity of the product of *any* two typologies is given by the diversity of the first, multiplied by the conditional diversity of the second (chain rule).

$$D(\tau_1, \tau_2) = D(\tau_1)\ D(\tau_2|\tau_1)$$

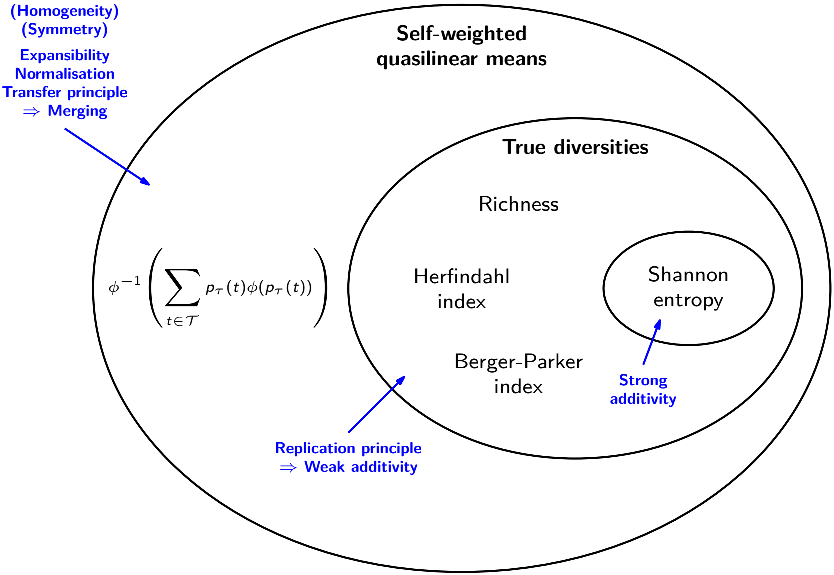
(τ_1, τ_2)	Comedy	Action	Drama	⋮	⋮
Novel	0	0	$\frac{4}{40}$	$\frac{1}{40}$	$\frac{2}{40}$
Movie	$\frac{2}{40}$	$\frac{2}{40}$	0	$\frac{1}{40}$	0
Comics	0	$\frac{6}{40}$	$\frac{3}{40}$	0	$\frac{3}{40}$
…	0	$\frac{2}{40}$	0	$\frac{1}{40}$	0
…	$\frac{2}{40}$	$\frac{4}{40}$	$\frac{3}{40}$	$\frac{3}{40}$	$\frac{1}{40}$

$\tau_2|\tau_1 = t_i$

0	0	$\frac{4}{7}$	$\frac{1}{7}$	$\frac{2}{7}$
$\frac{2}{5}$	$\frac{2}{5}$	0	$\frac{1}{5}$	0
0	$\frac{6}{12}$	$\frac{3}{12}$	0	$\frac{3}{12}$
0	$\frac{2}{3}$	0	$\frac{1}{3}$	0
$\frac{2}{13}$	$\frac{4}{13}$	$\frac{3}{13}$	$\frac{3}{13}$	$\frac{1}{13}$

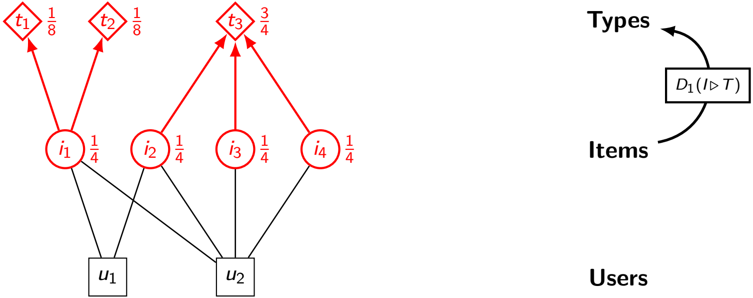
τ_1	$\frac{4}{40}$	$\frac{14}{40}$	$\frac{10}{40}$	$\frac{6}{40}$	$\frac{6}{40}$

Summary of diversity measures



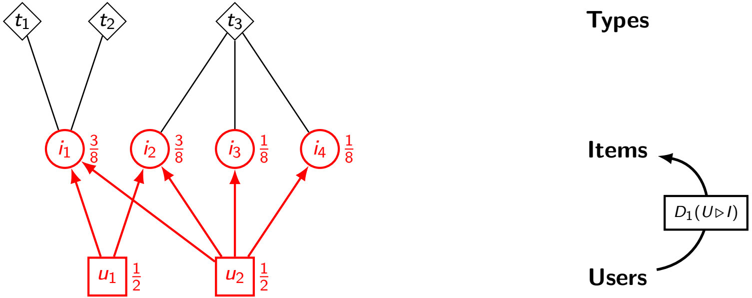
Application to tripartite graphs

Given a diversity measure D... (Here, we use Shannon entropy D_1.)



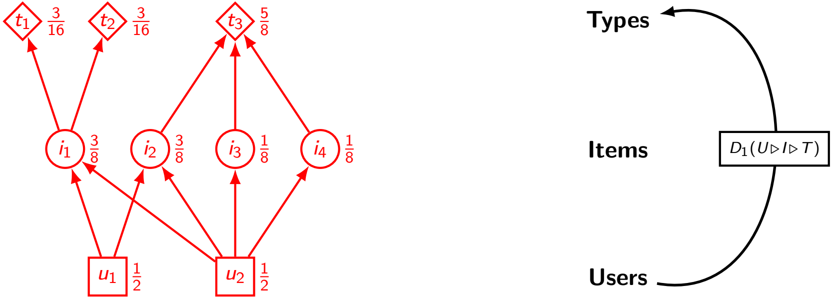
Application to tripartite graphs

Given a diversity measure D... (Here, we use Shannon entropy D_1.)



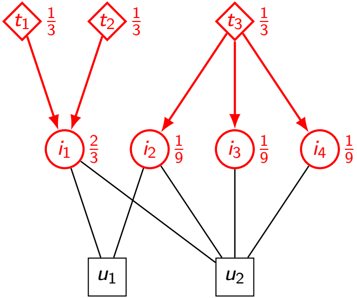
Application to tripartite graphs

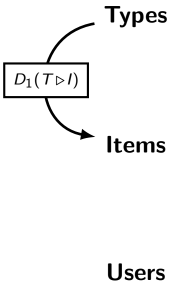
Given a diversity measure D... (Here, we use Shannon entropy D_1.)



Application to tripartite graphs

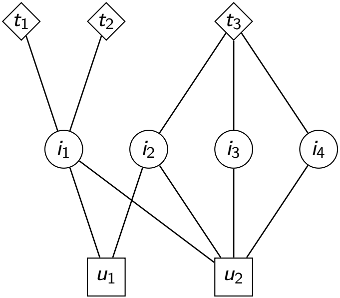
Given a diversity measure D... (Here, we use Shannon entropy D_1.)

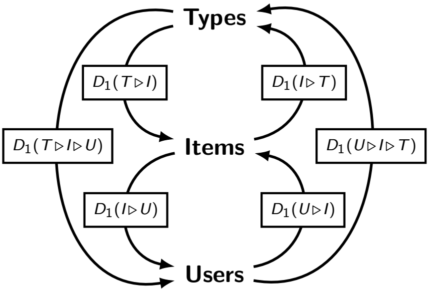




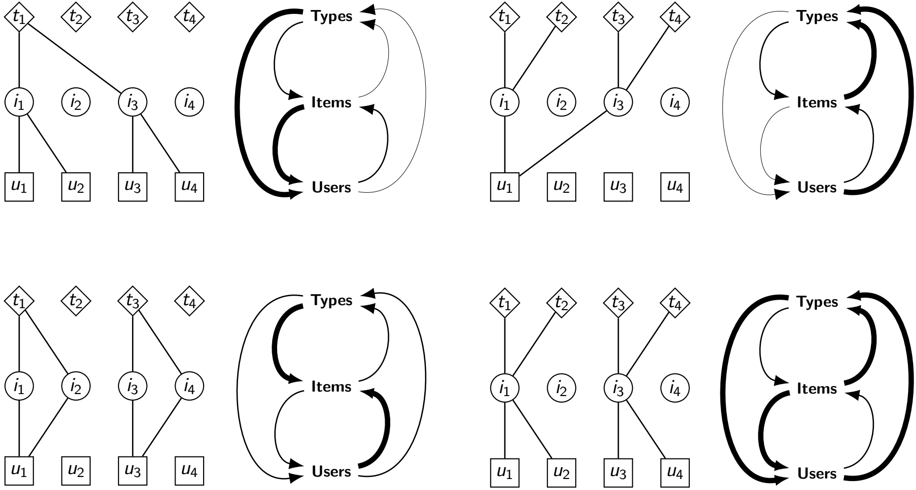
Application to tripartite graphs

Given a diversity measure D... (Here, we use Shannon entropy D_1.)

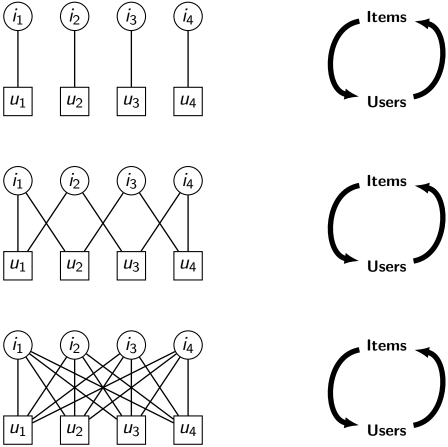




Application to tripartite graphs

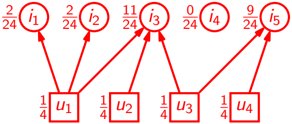


Individual Diversity vs. System Diversity (1/2)

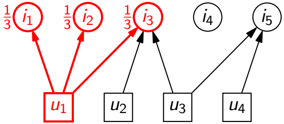


Same diversity (of items and of users)

Individual Diversity vs. System Diversity (2/2)



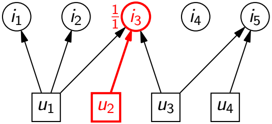
System diversity
$D_1(U \triangleright I) \approx 3.13$



Individual diversity of user u_1
$D_1(U \triangleright I \mid U = u_1) = 3$

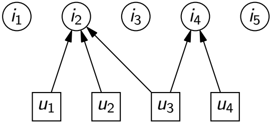
Mean individual diversity
$D_1(U \triangleright I \mid U) \approx 1.57$

$$D_1(U \triangleright I \mid U) = \prod_{u \in \mathcal{U}} D_1(U \triangleright I \mid U = u)^{\frac{1}{|\mathcal{U}|}}$$

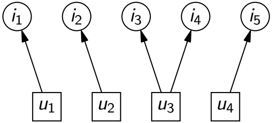


Individual diversity of user u_2
$D_1(U \triangleright I \mid U = u_2) = 1$

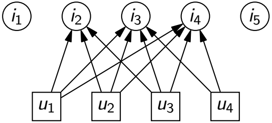
Individual Diversity vs. System Diversity



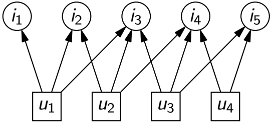
Weak individual diversity
Weak system diversity



Weak individual diversity
Strong system diversity

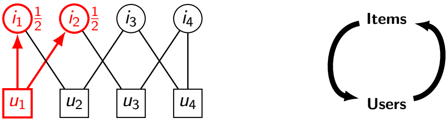


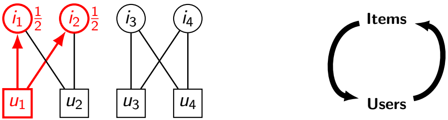
Strong individual diversity
Weak system diversity



Strong individual diversity
Strong system diversity

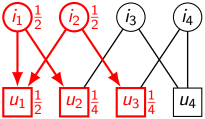
Diversity within communities

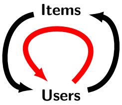


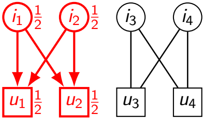


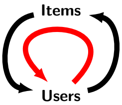
Same system diversity
Same individual diversity

Diversity within communities









Same system diversity
Same individual diversity

Different "retroactive" individual diversity!

The End

Thank you for your attention