# Biased Diffusion in Hyper-Bag-Graphs

CTW 2020
Online conference
16.09.2020

Xavier Ouvrard[1], Jean-Marie Le Goff[1], Stéphane Marchand-Maillet[2]

www.infos-informatique.net

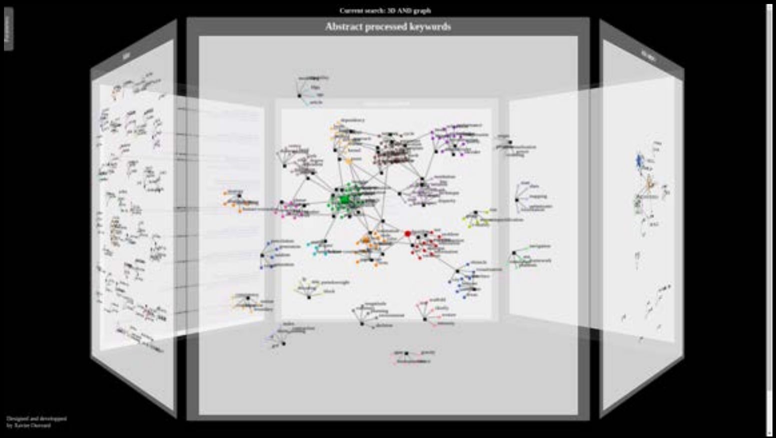
(1) CERN (2) UNIVERSITÉ DE GENÈVE

# Context (I): Information Retrieval


**With traditional verbatim browser:**

- The **output: linear** information
- To **refine** information: perform a **new search**
- **Complex query**: can be **hazardous**
- Accessing other facets of the information space => **perform different searches**

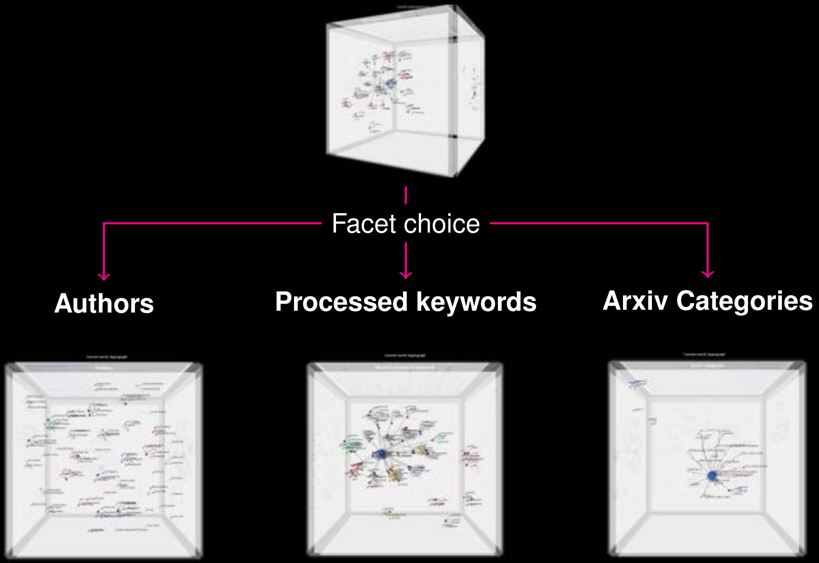
# Context (II): Information space




**But in fact:**

- A space of information is **multi-faceted**
- Much more information is **available** or **can be extracted**
- **Hb-graphs** highlight how the **data instances** are **linked** and allow **additional information** to be displayed

# Context: Facets of the Information Space




- **Information space** = interconnected networks of co-occurrences

# Multisets and Co-occurrences

**Multisets:**

**Multiset**: a universe and a multiplicity function $\mathfrak{A}_m = (A, m)$

**Natural multiset**: the range of the multiplicity function is a subset of $\mathbb{N}$.

In natural multisets: two views:

weighted set: $\mathfrak{A}_m = \left\{ x_1^{m_1}, \dots, x_n^{m_n} \right\}$

collection of objects $\mathfrak{A}_m = \left\{ \left\{ \underbrace{x_1, \dots, x_1}_{m_1 \text{ times}}, \dots, \underbrace{x_n, \dots, x_n}_{m_n \text{ times}} \right\} \right\}$

=> a **co-occurrence** appears as a multiset

=> in **literature**, network of co-occurrences approximated with pairwise relationships (graphs) or with the support of the multiset (hypergraphs)

# Hb-graphs

**Hb-graph $\mathcal{H} = (V, \mathfrak{E})$:** family of multisets $\mathfrak{E} = (\mathfrak{e}_i)_{i \in I}$, with $I = [\![p]\!]$- called **hb-edges** - where the hb-edges have:

- same universe $V = \{v_1, \ldots, v_n\}$, called **vertex set**.
- support a subset of $V$.
- each hb-edge has its own multiplicity function $m_{\mathfrak{e}} : V \to \mathbb{W}$ where $\mathbb{W} \subset \mathbb{R}^+$.

**Incidence matrix of hb-graphs:**

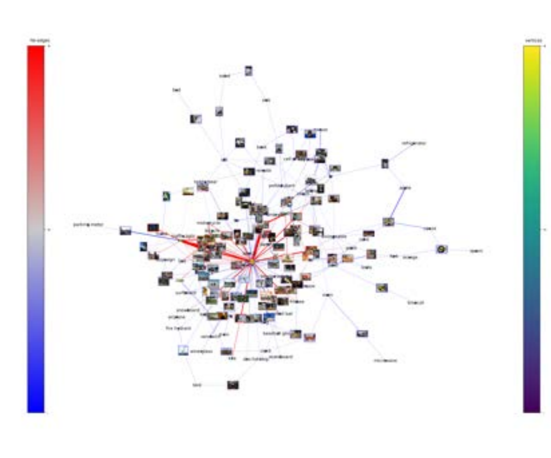
$$H = [m_j(v_i)]_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant p}}$$

**Different application of hb-graphs:**

- **Network of co-occurrences** are hb-graphs

=> hb-graph framework for modeling information space

# ML interest of Hb-graphs



**Exchange-based diffusion in hb-graphs:** Ouvrard et al. [2018, 2019]

- **Stochastic process**
- Allows **generalised random walk**
- Defines **a ranking** of vertices and hb-edges (akin to PageRank)
- Enables **coarsening** of hb-graphs and thus data landscape

=> Diffusion used for doing aggregation ranking (talk of last year @ CTW 2019)

# Why a biased exchange-based diffusion?

- In **standard exchange-based diffusion**:
high m-cardinality & high m-degree => highly ranked
- But depending on the focus of the search:
**different facets = different importance of the m-cardinality**

# Related work

- Dehmer and Mowshowitz [2011]: **abstract information function** $f : V \to \mathbb{R}^+$ such that for every: $v_i \in V$ :

$$p^f(v_i) = \frac{f(v_i)}{\sum_{j \in [\![ |V| ]\!]} f(v_j)}.$$

- Zlatić et al. [2010]: bias in the transition probability of a random walk in order to explore communities in a network

**Transition probability between vertex $v_i$ and $v_j$** given by:

$$T_{ij}(x, \beta) = \frac{a_{ij} e^{\beta x_i}}{\sum_l a_{lj} e^{\beta x_l}},$$

where $A = (a_{ij})_{i,j \in [\![ n ]\!]}$ is the adjacency matrix of the graph and $\beta$ is a parameter.

# Biased exchange-based diffusion in hb-graphs I

Considered: a weighted hb-graph $\mathfrak{H} = (V, \mathfrak{E}, w_e)$ with $V = \{v_i : i \in [\![n]\!]\}$ and $\mathfrak{E} = (\mathfrak{e}_j)_{j \in [\![p]\!]}$ ; we write $H = [m_{\mathfrak{e}_j}(v_i)]_{\substack{i \in [\![n]\!] \\ j \in [\![p]\!]}}$ the incidence matrix of the hb-graph.

**Vertex level**

1. Vertex abstract information function and corresponding probability

- **hb-edge based vertex abstract information function**: $f_V : V \times E \to \mathbb{R}^+$.
- **vertex abstract information function**: $F_V : V \to \mathbb{R}^+$ such that:

$F_V(v_i) \triangleq \sum_{j \in [\![p]\!]} f_V(v_i, \mathfrak{e}_j)$.

- **probability** corresponding to this hb-edge based vertex abstract information as: $p^{f_V}(\mathfrak{e}_j | v_i) \triangleq \frac{f_V(v_i, \mathfrak{e}_j)}{F_V(v_i)}$.

*For instance:* $f_V(v_i, \mathfrak{e}_j) = m_j(v_i) w(\mathfrak{e}_j)$ *and* $F_V(v_i) = d_{w,v_i}$

*=> retrieve the phase 1 of the exchange-based diffusion*

# Biased exchange-based diffusion in hb-graphs II

2. Now, we introduce a **bias on the abstract information**:
- **vertex bias function**: $g_V : \mathbb{R}^+ \to \mathbb{R}^+$ applied to $f_V(v_i, \mathfrak{e}_j)$
- **biased probability** on the transition from vertices to hb-edges defined as:

$$\widetilde{p_V}(\mathfrak{e}_j | v_i) \triangleq \frac{g_V(f_V(v_i, \mathfrak{e}_j))}{G_V(v_i)}$$

*In the exchange-based diffusion, we have used:* $g_V(x) = x$.
Typical choices for $g_V$ are: $g_V(x) = x^\alpha$ or $g_V(x) = e^{\alpha x}$. When $\alpha > 0$, higher values of $f_V$ are encouraged, and on the contrary, when $\alpha < 0$ smaller values of $f_V$ are encouraged.

# Biased exchange-based diffusion in hb-graphs III

**Hb-edge level**

1. Hb-edge abstract information function and corresponding probability

- **vertex-based hb-edge abstract information function**: $f_E : E \times V \rightarrow \mathbb{R}^+$.
- **hb-edge abstract information function** is defined as the function:

$F_E : V \rightarrow \mathbb{R}^+$, such that: $F_E(\mathfrak{e}_j) \triangleq \sum_{i \in [\![n]\!]} f_E(\mathfrak{e}_j, v_i)$.

- **probability** corresponding to the vertex-based hb-edge abstract information is defined as: $p^{f_E}(v_i | \mathfrak{e}_j) \triangleq \frac{f_E(\mathfrak{e}_j, v_i)}{F_E(\mathfrak{e}_j)}$.

# Biased exchange-based diffusion in hb-graphs IV

2. Now, we introduce a bias on the abstract information:

- **hb-edge bias function**: $g_E : \mathbb{R}^+ \to \mathbb{R}^+$ applied to $f_E(\mathfrak{e}_j, v_i)$,
- **hb-edge overall bias** defined as: $G_E(\mathfrak{e}_j) \triangleq \sum_{i \in [\![n]\!]} g_E(f_E(\mathfrak{e}_j, v_i))$.
- **biased probability** on the transition from hb-edges to vertices is defined as:

$$\widetilde{p_E}(v_i|\mathfrak{e}_j) \triangleq \frac{g_E(f_E(\mathfrak{e}_j, v_i))}{G_E(\mathfrak{e}_j)}$$

# Biased exchange-based diffusion in hb-graphs V

**Building a two-phase step diffusion by exchange:**

- Vertices hold an information value at time $t$ given by: $\alpha_t : V \to [0;1]$.
- Hb-edges hold an information value at time $t$ given by: $\epsilon_t : \mathfrak{E} \to [0;1]$.
- Information value of vertices: $I_t(V) = \sum\limits_{v_i \in V} \alpha_t(v_i)$
- Information value of hb-edges: $I_t(\mathfrak{E}) = \sum\limits_{\mathfrak{e}_j \in \mathfrak{E}} \epsilon_t(\mathfrak{e}_j)$
- Information value of the hb-graph: $I_t(\mathfrak{H}) = I_t(V) + I_t(\mathfrak{E})$.
- Closed non dissipative system:

The hb-graph information value is **kept constant** overtime to 1.

# Biased exchange-based diffusion in hb-graphs VI

- **Initialisation**: vertices have the information:

$\alpha_0 (v_i) = \alpha_{\text{ref}} = \frac{1}{|V|}$. Hence: $\mathfrak{e}_j \in \mathfrak{E}$, $\epsilon_0 (\mathfrak{e}_j) = 0$.

- **Two phases per time step**:
  - From $t$ to $t + \frac{1}{2}$ : **vertices distribute their values to hb-edges**:

$$\delta\epsilon_{t+\frac{1}{2}} (\mathfrak{e}_j | v_i) = \widetilde{p_V} (\mathfrak{e}_j | v_i) \alpha_t (v_i)$$

$$\epsilon_{t+\frac{1}{2}} (\mathfrak{e}_j) = \sum_{i=1}^{n} \delta\epsilon_{t+\frac{1}{2}} (\mathfrak{e}_j \mid v_i)$$

$$\alpha_{t+\frac{1}{2}} (v_i) = 0$$

  - From $t + \frac{1}{2}$ and $t + 1$: **hb-edges distribute their values to vertices**:

$$\delta\alpha_{t+1} (v_i \mid \mathfrak{e}_j) = \widetilde{p_E} (v_i | \mathfrak{e}_j) \epsilon_{t+\frac{1}{2}} (\mathfrak{e}_j) .$$

$\alpha_{t+1} (v_i) = \sum_{j=1}^{p} \delta\alpha_{t+1} (v_i \mid \mathfrak{e}_j) \epsilon_{t+1} (\mathfrak{e}_j) = 0.$

# Biased exchange-based diffusion in hb-graphs VII

To summarize (... details on Arxiv):

$$P_{\mathfrak{E},t+\frac{1}{2}} = P_{V,t} G_V^{-1} B_V. \tag{1}$$

$$P_{\mathfrak{E},t+\frac{1}{2}} G_{\mathfrak{E}}^{-1} B_E = P_{V,t+1}. \tag{2}$$

$$P_{V,t+1} = P_{V,t} G_V^{-1} B_V G_{\mathfrak{E}}^{-1} B_E. \tag{3}$$

- Writing $T = G_V^{-1} B_V G_{\mathfrak{E}}^{-1} B_E$, it follows from 3:

$$P_{V,t+1} = P_{V,t} T.$$

- $T$ is a **square row stochastic matrix** of dimension $n$.

Assuming that the hb-graph is connected, the biased feature exchange-based diffusion matrix $T$ is **aperiodic and irreducible**.
The fact that $T$ is a stochastic matrix aperiodic and irreducible for a connected hb-graph ensures that $(\alpha_t)_{t\in\mathbb{N}}$ **converges to a stationary state** which is the probability vector $\pi_V$ associated to the eigenvalue 1 of $T$.
**No explicit expression of the stationary state vector**

# Results and evaluation I

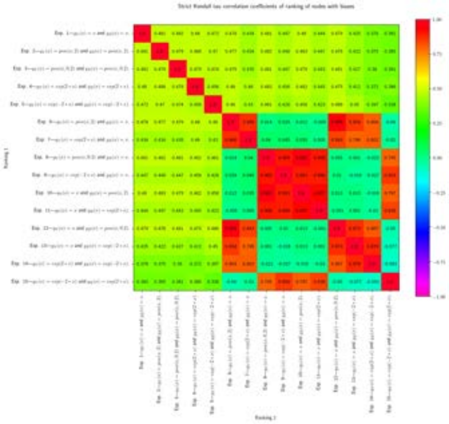
| Experiment | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Vertex bias function $g_V(x) =$ | $x$ | $x^2$ | $x^{0.2}$ | $e^{2x}$ | $e^{-2x}$ |
| Hb-edge bias function $g_E(x) =$ | $x$ | $x^2$ | $x^{0.2}$ | $e^{2x}$ | $e^{-2x}$ |

| Experiment | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|
| Vertex bias function $g_V(x) =$ | $x^2$ | $e^{2x}$ | $x^{0.2}$ | $e^{-2x}$ | $x$ |
| Hb-edge bias function $g_E(x) =$ | $x$ | $x$ | $x$ | $x$ | $x^2$ |

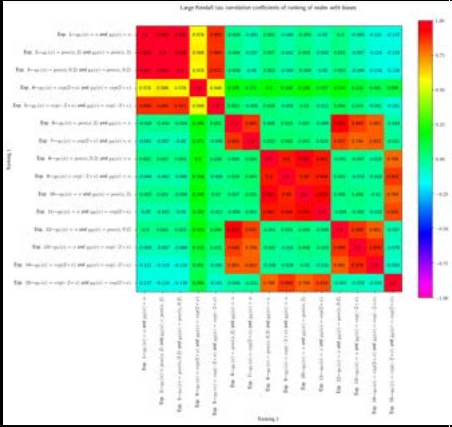
| Experiment | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|
| Vertex bias function $g_V(x) =$ | $x$ | $x$ | $x$ | $e^{2x}$ | $e^{-2x}$ |
| Hb-edge bias function $g_E(x) =$ | $e^{2x}$ | $x^{0.2}$ | $e^{-2x}$ | $e^{-2x}$ | $e^{2x}$ |

Biases used during the 15 experiments.

# Results and evaluation II



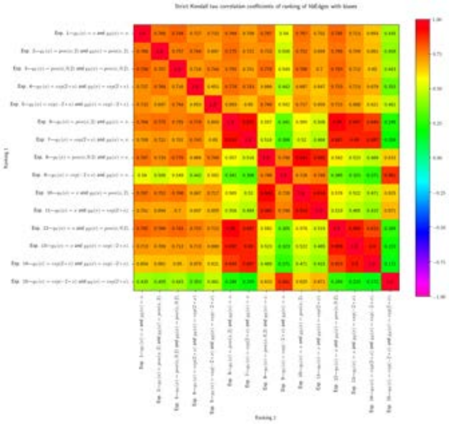
(a) Strict Kendall tau correlation coefficient



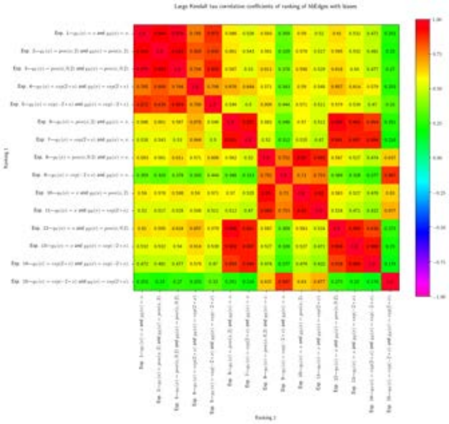
(b) Large Kendall tau correlation coefficient

for **node** ranking with biases. Realized on 100 random hb-graphs with 200 hb-edges of maximal size 20, with 5 groups.

# Results and evaluation III



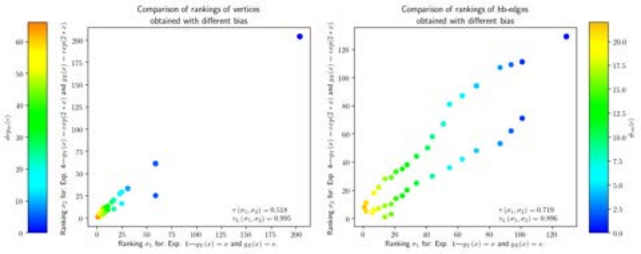
(a) Strict Kendall tau correlation coefficient



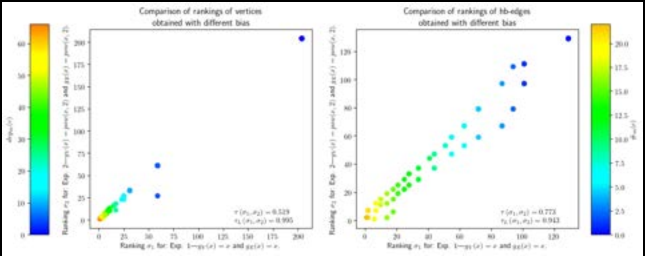
(b) Large Kendall tau correlation coefficient

for **hb-edge** ranking with biases. Realized on 100 random hb-graphs with 200 hb-edges of maximal size 20, with 5 groups.

# Results and evaluation IV

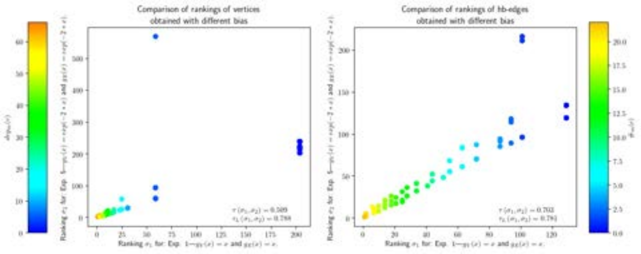



(a) First ranking: $g_V(x) = x$ and $g_E(x) = x$; Second ranking: $g_V(x) = e^{2x}$ and $g_E(x) = e^{2x}$.

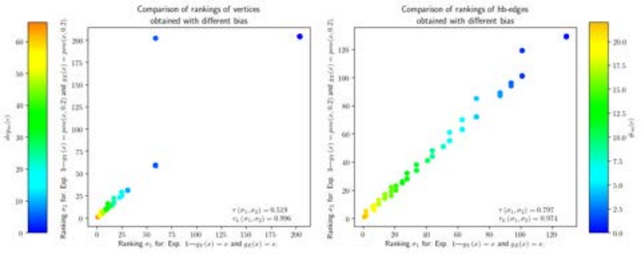



(b) First ranking: $g_V(x) = x$ and $g_E(x) = x$; Second ranking: $g_V(x) = x^2$ and $g_E(x) = x^2$.

# Results and evaluation V




(c) First ranking: $g_V(x) = x$ and $g_E(x) = x$; Second ranking: $g_V(x) = e^{-2x}$ and $g_E(x) = e^{-2x}$.




(d) First ranking: $g_V(x) = x$ and $g_E(x) = x$; Second ranking: $g_V(x) = x^{0.2}$ and $g_E(x) = x^{0.2}$.

# Conclusion & Future work

With these first results:

- There is an interest to apply different biases to explore differently the hb-graph => impact on hb-edges and nodes ranking
- **Tunable diffusion** to tune adequately the ranking of the facets

As FW:

Apply this approach to real cases:

- a publication database for refining queries
- an image case

# Thank you for your attention!

Collaboration spotting

*Leveraging insight into your data network by viewing co-occurrences while navigating across different perspectives.*

**More information:**

- http://collspotting.web.cern.ch
- https://www.infos-informatique.net
- xavier.ouvrard@cern.ch

# Bibliography I


Matthias Dehmer and Abbe Mowshowitz. A history of graph entropy measures. *Information Sciences*, 181(1):57–78, 2011.

Xavier Ouvrard, Jean-Marie Le Goff, and Stephane Marchand-Maillet. Diffusion by exchanges in hb-graphs: Highlighting complex relationships. *CBMI Proceedings*, 2018.

Xavier Ouvrard, Jean-Marie Le Goff, and Stephane Marchand-Maillet. Diffusion by exchanges in hb-graphs: Highlighting complex relationships extended version. *Arxiv:1809.00190v2*, 2019.

Vinko Zlatić, Andrea Gabrielli, and Guido Caldarelli. Topologically biased random walk and community finding in networks. *Physical Review E*, 82(6): 066109, 2010.