Tuning Ranking With Biased Exchange-Based Diffusion on Hyper-bag-graphs

Xavier Ouvrard[1,3], Jean-Marie Le Goff[1] and Stéphane Marchand-Maillet[2]

[1] CERN, Esplanade des Particules, 1, CH-1211 Meyrin (Switzerland)
[2] University of Geneva, CUI, Battelle (Bat A), CH-1227 Carouge (Switzerland)
[3] xavier.ouvrard@cern.ch

Abstract

Co-occurrence networks can be adequately modeled by hyper-bag-graphs (hb-graphs for short). A hb-graph is a family of multisets having same universe, called the vertex set. An efficient exchange-based diffusion scheme has been previously proposed that allows the ranking of both vertices and hb-edges. In this article, we extend this scheme to allow biases of different kinds and explore their effect on the different rankings obtained. The biases enhance the emphasize on some particular aspects of the network. The full text with proofs and results can be found in [1].
Keywords : *hyper-bag-graphs, biased diffusion, ranking*

1 Introduction, Background and Related Work

Co-occurrence networks have been shown to be modeled efficiently by using hyper-bag-graphs (hb-graphs for short) introduced in [2]. When considering an information space[1], different co-occurrence networks are interconnected using a common reference used for building the co-occurrences of different types [3]. Depending on the information the different co-occurrence networks carry, the ranking of the information held by the associated hb-graphs has to be performed on different features, and the importance stressed on the lower, higher, or medium values. Typically considering a publication information space, and considering the co-occurrences of organizations, countries and subjects in the publications, we might be interested on focusing on publications that have co-occurrence of small cardinality for subject categories and high cardinality in the number of organizations and a medium number of countries. Hence, the necessity of extending the exchange-based diffusion that is already coupled to a biased random walk given in [4] to a more general approach. The full text with proofs and results can be found in [1].

A **hb-graph** $\mathfrak{H} = (V, \mathfrak{E})$ is a family of multisets $\mathfrak{E} = (\mathfrak{e}_j)_{j\in[\![p]\!]}$ of same universe $V = \{v_i : i \in [\![n]\!]\}$[2]. The elements of $\mathfrak{E}$ are called the hb-edges; each hb-edge $\mathfrak{e}_j$, $j \in [\![p]\!]$, is a multiset of universe V and of multiplicity function: $m_{\mathfrak{e}_j} : V \to \mathbb{R}^+$. The m-cardinality $\#_m\mathfrak{e}_j$ of a hb-edge is: $\#_m\mathfrak{e}_j \triangleq \sum_{i\in[\![n]\!]} m_{\mathfrak{e}_j}(v_i)$. For more information on hb-graphs, the interested reader can refer to [5] for a full introduction. A weighted hb-graph has hb-edges having a weight given by: $w_e : \mathfrak{E} \to \mathbb{R}^+$.

Different approaches exist in the literature for studying networks based on graphs. In [6], the authors introduce an abstract information function which is associated to a probability for each vertex. In [7], a bias is introduced in the transition probability of a random walk in order to explore communities in a network.

[1] For a video showing, an Arxiv use case: https://www.infos-informatique.net
[2] $[\![n]\!]$ designates the integers between one 1 and n included.

2 Biased Diffusion in Hb-graphs

Let us consider 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.

2.1 Abstract Information Functions and Bias

A **hb-edge based vertex abstract information function:** $f_V : V \times \mathfrak{E} \to \mathbb{R}^+$ is considered. The exchange-based diffusion presented in [8, 4] is a particular example of biased diffusion. An unbiased diffusion would be to have a vertex abstract function and a hb-edge vertex function that is put to 1 for every vertices and hb-edges, i.e. equi-probability for every vertices and every hb-edges.

The **vertex abstract information function** is defined as the function: $F_V : V \to \mathbb{R}^+$ such that: $F_V(v_i) \triangleq \sum_{j\in[\![p]\!]} f_V(v_i, \mathfrak{e}_j)$. The **probability corresponding to this hb-edge based vertex abstract information** is defined as: $p^{f_V}(\mathfrak{e}_j|v_i) \triangleq \dfrac{f_V(v_i,\mathfrak{e}_j)}{F_V(v_i)}$. Considering a **vertex bias function**: $g_V : \mathbb{R}^+ \to \mathbb{R}^+$ applied to $f_V(v_i, \mathfrak{e}_j)$, we can define a **biased probability on the transition from vertices to hb-edges** as:

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

where $G_V(v_i)$, the **vertex overall bias**, is defined as: $G_V(v_i) \triangleq \sum_{j\in[\![p]\!]} g_V(f_V(v_i,\mathfrak{e}_j))$. 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.

Similarly, the **vertex-based hb-edge abstract information function** is defined as: $f_{\mathfrak{E}} : \mathfrak{E} \times V \to \mathbb{R}^+$. The **hb-edge abstract information function**: $F_{\mathfrak{E}} : V \to \mathbb{R}^+$ is defined as: $F_{\mathfrak{E}}(\mathfrak{e}_j) \triangleq \sum_{i\in[\![n]\!]} f_{\mathfrak{E}}(\mathfrak{e}_j, v_i)$. The **probability corresponding to the vertex-based hb-edge abstract information** is defined as: $p^{f_{\mathfrak{E}}}(v_i|\mathfrak{e}_j) \triangleq \dfrac{f_{\mathfrak{E}}(\mathfrak{e}_j,v_i)}{F_{\mathfrak{E}}(\mathfrak{e}_j)}$. If we now consider a vertex bias function: $g_{\mathfrak{E}} : \mathbb{R}^+ \to \mathbb{R}^+$ applied to $f_{\mathfrak{E}}(\mathfrak{e}_j, v_i)$, we can define a **biased probability on the transition from hb-edges to vertices** as:

$$\widetilde{p_{\mathfrak{E}}}(v_i|\mathfrak{e}_j) \triangleq \frac{g_{\mathfrak{E}}(f_{\mathfrak{E}}(\mathfrak{e}_j,v_i))}{G_{\mathfrak{E}}(\mathfrak{e}_j)},$$

where $G_{\mathfrak{E}}(\mathfrak{e}_j)$, the hb-edge overall bias is defined as: $G_{\mathfrak{E}}(\mathfrak{e}_j) \triangleq \sum_{i\in[\![n]\!]} g_{\mathfrak{E}}(f_{\mathfrak{E}}(\mathfrak{e}_j,v_i))$. Typical choices for $g_{\mathfrak{E}}$ are: $g_{\mathfrak{E}}(x) = x^\alpha$ or $g_{\mathfrak{E}}(x) = e^{\alpha x}$. When $\alpha > 0$, higher values of $f_{\mathfrak{E}}$ are encouraged, and on the contrary, when $\alpha < 0$ smaller values of $f_{\mathfrak{E}}$ are encouraged.

2.2 Biased Diffusion by Exchange

A two-phase step diffusion by exchange is now considered—with a similar approach to [8, 4]—taking into account the biased probabilities on vertices and hb-edges. At time t, the vertices hold an information value given by: $\alpha_t : V \to [0;1]$ and the hb-edges via: $\epsilon_t : \mathfrak{E} \to [0;1]$.

We write $P_{V,t} = (\alpha_t(v_i))_{i\in[\![n]\!]}$ the row state vector of the vertices at time t and $P_{\mathfrak{E},t} = (\epsilon_t(\mathfrak{e}_j))_{j\in[\![p]\!]}$ the row state vector of the hb-edges. We call information value of the vertices, the value: $I_t(V) \triangleq \sum_{v_i\in V} \alpha_t(v_i)$ and $I_t(\mathfrak{E}) \triangleq \sum_{\mathfrak{e}_j\in\mathfrak{E}} \epsilon_t(\mathfrak{e}_j)$ the one of the hb-edges. We write: $I_t(\mathfrak{H}) \triangleq I_t(V) + I_t(\mathfrak{E})$.

The initialisation is done such that $I_0(\mathfrak{H}) = 1$. At the diffusion process start, the vertices concentrate uniformly and exclusively all the information value. Writing $\alpha_{\text{ref}} = \frac{1}{|V|}$, we set for all $v_i \in V : \alpha_0(v_i) = \alpha_{\text{ref}}$ and for all $\mathfrak{e}_j \in \mathfrak{E}$, $\epsilon_0(\mathfrak{e}_j) = 0$.

At every time step, the first phase starts at time t and ends at $t + \frac{1}{2}$, where values held by the vertices are shared completely to the hb-edges, followed by the second phase between time $t + \frac{1}{2}$ and $t + 1$, where the exchanges take place the other way round. The exchanges between vertices and hb-edges aim at being conservative on the global value of α_t and ϵ_t distributed over the hb-graph.

During the first phase between time t and time $t + \frac{1}{2}$, the contribution to the value $\epsilon_{t+\frac{1}{2}}(\mathfrak{e}_j)$ from the vertex v_i is given by:

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

and:

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

We have:

$$\alpha_{t+\frac{1}{2}}(v_i) = \alpha_t(v_i) - \sum_{j=1}^{p} \delta\epsilon_{t+\frac{1}{2}}(\mathfrak{e}_j \mid v_i).$$

It holds: $\forall i \in [\![n]\!] : \alpha_{t+\frac{1}{2}}(v_i) = 0$, and: $I_{t+\frac{1}{2}}(\mathfrak{H}) = 1$.

We introduce the **vertex overall bias matrix**: $G_V \overset{\Delta}{=} \text{diag}\left((G_V(v_i))_{i\in[\![n]\!]}\right)$ and the **biased vertex-feature matrix**: $B_V \overset{\Delta}{=} [g_V(f_V(v_i, \mathfrak{e}_j))]_{\substack{i\in[\![n]\!] \\ j\in[\![p]\!]}}$. It holds:

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

During the second phase that starts at time $t + \frac{1}{2}$, the values held by the hb-edges are transferred to the vertices. The contribution to $\alpha_{t+1}(v_i)$ given by a hb-edge $\mathfrak{e}_j$ is proportional to $\epsilon_{t+\frac{1}{2}}$ in a factor corresponding to the biased probability $\widetilde{p_{\mathfrak{E}}}(v_i|\mathfrak{e}_j)$:

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

Hence, we have: $\alpha_{t+1}(v_i) = \sum_{j=1}^{p} \delta\alpha_{t+1}(v_i \mid \mathfrak{e}_j)$ and:

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

It holds: $\forall j \in [\![p]\!] : \epsilon_{t+1}(\mathfrak{e}_j) = 0$, and: $I_{t+1}(\mathfrak{H}) = 1$.

We now introduce $G_{\mathfrak{E}} \overset{\Delta}{=} \text{diag}\left((G_{\mathfrak{E}}(\mathfrak{e}_j))_{j\in[\![p]\!]}\right)$ the diagonal matrix of size $p \times p$ and the **biased hb-edge-feature matrix**: $B_{\mathfrak{E}} \overset{\Delta}{=} [g_{\mathfrak{E}}(f_{\mathfrak{E}}(\mathfrak{e}_j, v_i))]_{\substack{j\in[\![p]\!] \\ i\in[\![n]\!]}}$, it comes:

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

Regrouping (1) and (2):

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

It is valuable to keep a trace of the intermediate state: $P_{\mathfrak{E},t+\frac{1}{2}} = P_{V,t} G_V^{-1} B_V$ as it records the information on hb-edges.

Writing $T = G_V^{-1} B_V G_{\mathfrak{E}}^{-1} B_{\mathfrak{E}}$, 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. Hence, $(\alpha_t)_{t\in\mathbb{N}}$

converges to a stationary state which is the probability vector π_V associated to the eigenvalue 1 of T. Nonetheless, due to the presence of the different functions for vertices and hb-edges, the simplifications do not occur anymore as in [8, 4] and thus we do not have an explicit expression for the stationary state vector of the vertices. The same occurs for the expression of the hb-edge stationary state vector $\pi_{\mathfrak{E}}$ which is still calculated from π_V using the following formula: $\pi_{\mathfrak{E}} = \pi_V G_V^{-1} B_V$.

3 Evaluation and Further Comments

We have randomly generated connected hb-graphs with 200 collaborations—built out of 10,000 potential vertices—with a maximum m-cardinality of 20, such that the hb-graph has five groups that are generated with two of the vertices chosen out of a group of 10, that have to occur in each of the collaboration; there are 20 vertices that have to stand as central vertices, i.e. that ensures the connectivity in between the different groups of the hb-graph.

The approach is similar to the one taken in [8, 4], using the same hb-edge based vertex abstract information function and the same vertex-based hb-edge abstract information function, but putting different biases. We compare the rankings obtained on vertices and hb-edges after 200 iterations of the exchange-based diffusion using the strict and large Kendall tau correlation coefficients for the different biases. We present the results as a visualisation of correlation matrices in [1].

The results obtained on randomly generated hb-graphs have still to be applied to real hb-graphs, with the known difficulty of the connectedness: it will be addressed in future work.

References

[1] X. Ouvrard, J.-M. L. Goff, and S. Marchand-Maillet, "Tuning ranking in co-occurrence networks with general biased exchange-based diffusion on hyper-bag-graphs," 2020. arXiv: 2003.07323 [cs.SI].

[2] X. Ouvrard, J.-M. Le Goff, and S. Marchand-Maillet, "Adjacency and Tensor Representation in General Hypergraphs. Part 2: Multisets, Hb-graphs and Related e-adjacency Tensors," *arXiv preprint arXiv:1805.11952*, 2018.

[3] X. Ouvrard, J.-M. Le Goff, and S. Marchand-Maillet, "The HyperBagGraph DataEdron: An Enriched Browsing Experience of Datasets," *LNCS, 46th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2020)*, 2020.

[4] X. Ouvrard, J.-M. Le Goff, and S. Marchand-Maillet, "Exchange-Based Diffusion in Hb-Graphs: Highlighting Complex Relationships in Multimedia Collections," *Multimedia Tools and Applications*.

[5] X. Ouvrard, J.-M. Le Goff, and S. Marchand-Maillet, "On Hb-graphs and their Application to General Hypergraph e-adjacency Tensor," *MCCCC32 Special Volume of the Journal of Combinatorial Mathematics and Combinatorial Computing, to be published*, 2019.

[6] M. Dehmer and A. Mowshowitz, "A history of graph entropy measures," *Information Sciences*, vol. 181, pp. 57–78, Jan. 2011.

[7] V. Zlatić, A. Gabrielli, and G. Caldarelli, "Topologically biased random walk and community finding in networks," *Physical Review E*, vol. 82, p. 066109, Dec. 2010.

[8] X. Ouvrard, J.-M. Le Goff, and S. Marchand-Maillet, "Diffusion by Exchanges in HB-Graphs: Highlighting Complex Relationships," *CBMI Proceedings*, 2018.