EXPLOITING INTRA-TYPE INFORMATION IN BIPARTITE COMMUNITY DETECTION

Paola Pesántez-Cabrera, Ananth Kalyanaraman, Mahantesh Halappanavar

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary: Classical bipartite community methods only take into account inter-type edge information—i.e., edges between vertices of two different types. We present a new form of bipartite modularity (as an objective function for community detection) that can enable methods to incorporate both intra-type and inter-type edge information. Preliminary results evaluating this new form are presented.

Introduction

Bipartite graphs serve as an effective way to represent the interplay between two different data types—e.g., gene vs. disease, plant vs. pollinator, etc. (e.g., [3, 8]). Here, vertices represent the individual entities of a data type, and edges represent the interaction between the entities of two different data types. The problem of community detection, when applied to such bipartite networks, is one of co-clustering the entities of the two different types based on their inter-type interactions.

However, in many applications, we may also have *intra-type* information, which may be critical in determining the co-clustering structures [6]. For instance, considering the sequence-based similarity between genes (intra-type) could either provide the additional basis for clustering a group of genes with a group of diseases or help reveal hidden links between disease groups.

Current methods for bipartite community detection are ill-equipped to handle such intra-type information when made available. More specifically, the *modularity* metrics that they use, to measure the goodness of clustering, use only inter-type edge information. Note that a naïve way to handle both inter- and intra-type information is to simply treat the graph as a general graph and run methods that are designed for general graphs. However, intra-type may or may not carry the same weight as inter-type; furthermore, the connectivity characteristics (e.g., sparsity of edges, degree distribution) could differ between inter- and intra-type edges.

Contribution: In this paper, we present a definition of bipartite modularity that would enable bipartite community detection methods to compute clustering structures taking into account *both* inter- and intra-type edges. Our definition extends the `Murata+` definition of [7].

Notation and Definitions: Let $G = (V_1 \cup V_2, E)$ denote a bipartite graph, where V_1 and V_2 represent vertices of two different types, and an edge $e_{ij} \in E$ represents a pairwise relationship between $i \in V_1$ and $j \in V_2$. M denotes the sum of the weights of all edges in E. We define an *augmented bipartite graph* as a bipartite graph which also allows edges between vertices of the same type—i.e., $G(V_1 \cup V_2, E \cup E')$, where every edge $e_{ij} \in E'$ is such that either $i, j \in V_1$ or $i, j \in V_2$.

The goal of bipartite community detection is to partition V_1 and V_2 into a set of communities such that the members of a community are highly "related" to one another than to the rest of the network. The degree of relatedness is typically captured in the modularity of clustering.

Classical Definitions of Bipartite Modularity

Multiple bipartite modularity definitions have been proposed [1, 4, 5, 7]. However, all the above definitions focus on establishing community structures only based on inter-type information. Guimerà *et al.* [4] focuses on connectivity from the perspective of only one vertex type. Barber [1] assumes and enforces a one-to-one correspondence between the communities from the different vertex types, whereas Murata's definition [5] overcomes this limitation. During analysis, we encountered an inconsistency in Murata's definition and proposed a variant called `Murata+` defined as follows [7]:

$$Q_B = \sum_C (\mathcal{E}_{C,\psi(C)} - \mathcal{A}_C \times \mathcal{A}_{\psi(C)}) + \sum_D (\mathcal{E}_{D,\psi(D)} - \mathcal{A}_D \times \mathcal{A}_{\psi(D)}) \quad (1)$$

Here, C and D represent a community in V_1 and V_2 respectively; $\psi(C)$ denotes a D that is identified as the *co-cluster mate* of C in V_2 (similar definition for $\psi(D)$); $\mathcal{E}_{C,\psi(C)}$ represents the fraction of inter-type edges from C to $\psi(C)$ (similar for $\mathcal{E}_{D,\psi(D)}$); and A_C (or A_D) denotes the fraction of edges contributed by community C (or D).

Proposed Definition of Bipartite Modularity

Given an augmented bipartite graph $G(V_1 \cup V_2, E \cup E')$, we assume (without loss of generality) that all edges have normalized weights. First, we define a positive weight $\alpha \in \mathbb{R}$ for using inter-type edges (implying, $1-\alpha$ for intra-type edges). We use $s(i,j)$ to denote the“similarity” score between vertices i and j of the same type. Let us consider the bipartite network formed by genes and drugs; then, $s_g(i,j)$ is a sequence-based similarity score between two genes i and j, while $s_d(i,j)$ is a structure-based similarity score between two drugs i and j. Based on the s function, we define β and ϕ factors for community C of genes as follows (for $i \neq j$):

$$\beta(C) = \frac{\sum\limits_{i,j \in C} s_g(i,j)}{\sum\limits_{i,j \in V_1} s_g(i,j)}, \quad \phi(C) = \frac{\sum_{i \in C, j \in V_1} s_g(i,j)}{\sum_{i,j \in V_1} s_g(i,j)}$$

Intuitively, $\beta(C)$ represents the relative intra-cluster similarity based solely on intra-type edges, whereas $\phi(C)$ is the fraction of intra-type edges (in V_1) contributed by community C.

Subsequently, we define the augmented variant of the `Murata+` modularity definition as follows:

$$\boxed{Q_B = \sum_{C} (\mathcal{E}'_C - \mathcal{A}'_C) + \sum_{D} (\mathcal{E}'_D - \mathcal{A}'_D)} \tag{2}$$

where:

$$\begin{aligned} \mathcal{E}'_C &= [\alpha\ \mathcal{E}_{C,\psi(C)}] + [(1-\alpha)\beta(C)] \\ \mathcal{A}'_C &= [\alpha\ \mathcal{A}_C \mathcal{A}_{\psi(C)}] + [(1-\alpha)(\phi(C)\phi(C))] \end{aligned}$$

Implementation: We implemented the proposed modularity into our `biLouvain` community detection tool [7] (https://github.com/paolapesantez/biLouvain). We use a multi-level iterative scheme where vertices determine their communities at each step. We implemented two variants of how a vertex chooses its destination community C_j from its current community C_i:

Strongly constrained (*SC*): C_j that maximizes the modularity gain ΔQ_B and $\beta(C_j)$, such that $\beta(C_j) \geq \lambda_{V_k}$ and $\beta(C_j) > \beta(C_i)$;

Weakly constrained (*WC*): C_j that maximizes ΔQ_B while $\beta(C_j) \geq \lambda_{V_k}$, where λ_{V_k} is a predetermined cutoff.

Experimental Results

Test data: We experimented with an Enzyme-Interaction binary bipartite network [2] that has 1,109 nodes (664 targets and 445 drugs) and 317,841 edges (2,926 inter-type, 220,116 targets intra-type, and 94,799 drugs intra-type). We use $\lambda_{V_1} = 0.03$ and $\lambda_{V_2} = 0.25$ obtained from the average similarity scores and based on [6].

Table 1: Evaluation on an Enzyme-Interaction data set.

Alpha	**Modularity Q_B**		**Correlation Coefficient(%)**		
α	**SC**	**WC**	α **comparison**	**SC**	**WC**
0.0	2.90E-05	2.90E-05	0.0 vs. 0.1	15.88	14.76
0.2	0.259	0.281	0.2 vs. 0.3	83.44	89.34
0.4	0.381	0.420	0.4 vs. 0.5	88.39	89.41
0.6	0.509	0.562	0.6 vs. 0.7	81.19	93.32
0.8	0.647	0.718	0.8 vs. 0.9	94.08	95.41
1.0	0.869	0.869	0.9 vs. 1.0	87.31	96.76

Table 1 shows how adding intra-type information impacts the final modularity. When $\alpha = 0.0$, Q_B is small because targets and drugs form a few large communities; contrarily to when $\alpha = 1.0$. When α is increased, Q_B also increases. The *SC* case provides a better run-time because being more restrictive reduces the amount of work needed. Correlation coefficient percentages show the degree of conservation in the clusters obtained across different α values. α values between 0.4 and 0.6 produce approximately consistent community outputs, implying that giving roughly equal weight to inter- and intra-type edges for this input data set is desirable. When comparing clusters for $\alpha = 0.0$ vs. $\alpha = 0.1$, the major difference is a consequence of inter-type information exclusion. Finally, the less restrictive the constraint, the better correlation between clusters.

References

[1] M. J. Barber. Modularity and community detection in bipartite networks. Physical Review E, 76(6):066102, 2007.

[2] B. Chen, Y. Ding, and D. J. Wild. Assessing Drug Target Association Using Semantic Linked Data. PLOS Computational Biology, 8(7):e1002574, 2012.

[3] M. Griffith, O. L. Griffith, A. C. Coffman, J. V. Weible, J. F. McMichael, N. C. Spies, J. Koval, et al. DGIdb: mining the druggable genome. Nature Methods, 10(12):1209–1210, 2013.

[4] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. Physical Review E, 76(3):036102, 2007.

[5] T. Murata. Detecting communities from bipartite networks based on bipartite modularities. In International Conference on Computational Science and Engineering, volume 4, pages 50–57. IEEE, 2009.

[6] G. Palma, M.-E. Vidal, and L. Raschid. Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning. In International Semantic Web Conference, pages 131–146. Springer, Cham, 2014.

[7] P. Pesántez-Cabrera and A. Kalyanaraman. Efficient Detection of Communities in Biological Bipartite Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, PP(99), 2017.

[8] J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015:bav028, 2015.