Communities in directed networks

Towards a hybrid model of semantic communities detection

Gamgne Domgue Félicité, Tsopze Norbert, Ndoundam René, Arnaud S. R. M. Ahouandjinou

IRD UMI 209 UMMISCO,
University of Yaounde I
Yaounde,Cameroon
felice.gamgne@gmail.com, fgamgne@uy1.uninet.cm
tsopze@uy1.uninet.cm
ndoundam@gmail.com
arnaud.ahouandjinou@univ-littoral.fr, ahou.arn@gmail.com
Laboratoire LISIC, Université du Littoral de la Cote d'Opale(ULCO),62228 Calais, France,

ABSTRACT. Community detection in directed network is of vital importance to find cohesive sub-groups. Many existing graph clustering methods mainly focus on the relational structure and vertex properties, but ignore edge directionality during the clustering task, in case of directed graphs. In this paper we propose a hybrid semantic similarity which includes node attribute informations along with the network structure and link semantic. Then by application of a partitioning clustering technique, we evaluate its performance and results on a built textual based dataset with ground truth. We argue that, depending on the kind of data we have and the type of results we want, the choice of the clustering method is important and we present some concrete examples for underlining this.

RÉSUMÉ. La détection des clusters orientés constitue davantage un challenge dans l'analyse des réseaux. Plusieurs approches de clustering s'attardent uniquement sur la structure et les attributs des noeuds, mais ignorent la sémantique portée par les liens dans le cas des graphes orientés. Dans cet article nous proposons une mesure de similarité hybride qui combine les informations structurelles, les attributs et l'orientation des liens. Par application de cette mesure à un algorithme de clustering , nous évaluons les performances de cette nouvelle approche sur un jeu de données que nous avons construit avec vérité de terrain. Selon le type de données exploité et le type de résultats escomptés, nous montrons que le choix de la méthode de classification est important via quelques illustrations.

KEYWORDS : Directed attributed network, Graph clustering, Link semantic, Social network

MOTS-CLÉS : Réseau orienté attribué, Clustering, Sémantique des liens, Réseau social

1. Introduction

Cluster extraction is one of the main tasks of descriptive modelisation in datamining area. Like this, most of graph partitioning methods, useful for strongly connected community detection [7], focus on relational structure, but ignore node properties or attributes. More the recent approaches tended to find cohesive subgroups by combining node attributes with link informations in graph. These informations only concerned the structure data like frequent link-pattern(neighbourhood and leadership). Nevertheless, combining these different data types leads to the problem of semantic classification, because of the "inconsistent" similarity measures omitting the link *semantic* (meaning edge's directionality). A new challenge in community detection consists on meaningful cluster extraction based on three parameters : structure, node attributes and link semantic. In this paper, we propose an hybrid technique dealing with the *semantic based topological* structure of the graph, and we show that with textual attributes joined to vertices, it is possible to extract semantic clusters. We perform our experiments through the construction of an attributed directed network with ground truth, Normalized Mutual Information (NMI) and Density measures are used for evaluations. The work of incorporating structural *semantic* and attribute data has not yet been throughout studied in the context of large social graphs. This is the motivation of our work for which key contributions are summarized next : studying of the relationship between semantic similarity of species in a food web network and showing that the type of data determine the result, thus a textual attribute strengthens the semantic topology and helps to discover more relevant communities.

The document is organized as follows. The Section 2 presents related works based on graphs partitioning methods that take into account both features and structure relationship. The formal description of the idea is presented in Section 3, then some hybrid approaches based on both links and attribute information are suggested in Section 4. An experimental study describing the constructed dataset and the expected results according to the technique are presented in the section 5. After that experiment description, an evaluation on different semi-hybrid and hybrid models are shown in the Section 6, and the Section 7 concludes the study.

2. Related works

The well-known graph clustering techniques use the relationships between vertices to partition the graph into several densely connected components, but do not use the properties of the nodes. The problem is to combine both graph data and attribute data simultaneously in order to detect clusters that are densely connected and similar in the attribute space. Few recent studies have addressed the problem of clustering in attributed networks. Next, we present a classification of the existing methods of clustering in attributed graph based on their methodological principles.

Edge weighting based approaches : In order to integrate the attribute or structure information in the clustering process, these methods define a node attribute similarity that will be used to weight the existing edges. In literature, some relevant approaches have been proposed [1]. The first approach of the following section is based on this idea.

Pattern-based approaches : These methods focus on the structure or relational property of the graph, based on kernels information Li et al. [2]. In the same way, Gamgne et al. [8] extracted kernels through the neighbourhood overlap. The relationship information

is based on either the structural equivalence i.e. two vertices belong to the same cluster if they own the same neighbours or leadership i.e. vertices are connected to the same leader. They defined a *kernel degree* measure which denotes the similarity of nodes in their roles of leader (high in-degree) or follower (low in-degree) as studied by Gamgne et al. [9]. Its limit is that it does not deal with node attributes.

Quality function optimization based approaches : This family of approaches extend the well-know graph based clustering methods to consider both attribute information and topological structure. Authors in [6] proposed an extension of the Louvain algorithm with a modification of modularity by including an attribute similarity metric. [5] propose the **I-Louvain** algorithm which uses the inertia based modularity combined with the Newman's modularity.

Unified distance based approaches : They consist in transforming the topological information of the network into a similarity or a distance function between vertices. Zhou et al. [4] exploit the attributes in order to extend the original graph to an augmented one. A graph partitioning is then carried out on this new augmented graph. A neighborhood random walk model is used to measure the node closeness on the augmented graph. Then, they proposed a **SA-Cluster** algorithm that make use of a random walk distance measure and K-Medoids approach for the measurement of a node's closeness.

All of these methods have the limit that their topological property does not deal with link semantic, meaning edge directionality in directed networks. Yet the majority of real-life networks are represented as directed graphs, and link direction helps in improving partition quality.

We present in the Section 4, methods handling both topological and node attributes and that are easy to use, while the next section shows how formally a generic clustering approach could be implemented.

3. Problem Statement

An attributed graph is denoted as $G = (V, E, W)$, where V is the set of nodes, E is set of edges, and W is the set of attributes associated to the nodes in V for describing their features. Each vertex v_i is described by a real attribute vector $d_i = (w_1(v_i), ..., w_j(v_i), ..., w_m(v_i))$ where $w_j(v_i)$ is the attribute value of vertex v_i on attribute w_j. Into such network, clustering of attributed graph should take into account both structure network and attribute information by achieving a good balance between the following two properties : *(i)* vertices within one cluster are closed to each other in terms of "structure", meaning that vertices are arranged according to a semantic pattern, while vertices between clusters are not patterned; *(ii)* vertices within one cluster are more similar by their attributes than vertices from different clusters that could have quite different attribute values. In this work, we consider that the partitioning process focuses both on *semantic based topology* and node attributes. In others words, the structure concept includes not only link density, but also link semantic. The approach consists in dividing the set of nodes V into a partition of k clusters C_i, such that :

1) $C_i \cap C_j \neq \Phi \; \forall i \neq j$ and $\cup_i C_i = |V|$, where Φ is an empty set,

2) The semantic similarity takes into account three criteria : the link density, the node attribute and the link direction,

3) Vertices within clusters are semantically connected, while the vertices in different clusters are sparsely connected.

Likewise, we assume that an information network like a food web network can be represented by an attributed directed graph. Then, species relationship corresponds to a network in which each vertex represents a species and is described by a vector $d_i = (w_{i1}, w_{i2})$ where w_{i1} is the discrete attribute according to the diet mode (0 for "$carnivorous$" and 1 for "$herbivorous$") and w_{i2} the textual attribute denoting mode of reproduction (either " $oviparous$" or "$viviparous$") ; an edge from node a to node b means that species a is consumed by species b (*"Prey-Predator"* relationship). Thus, partitioning this kind of graph leads to integrate both ($density$ and $semantic$) topological and ($discrete$ or $textual$) attribute knowledge.

4. Clustering Graph models

Approaches for graph clustering described in this section separately handle both relational information and vertex attributes, and differ by their manner of combining relational data and attributes.

4.1. Attribute and Relational based clustering methods

Attribute based clustering method first exploits attributes by graph enrichment through a node attribute similarity (NAS) function [1,4,6]. According to the **SA-Cluster** method [4], the unified random walk distance is applied to an augmented graph. On the other hand, cosine distance between vertices v_i and v_j could be used, as defined as $SimA(v_i, v_j)$ in **SAC1** method [6].

In the relational based clustering model, structural properties are considered first through either a neighbourhood similarity. Li in [2] proposed a hierarchical clustering by filtering process of cores (kernels) based on structural information, then merging them by their attributes similarity. The core filtering is based on a frequent itemsets process through a similarity we labelled here $simS(v_i, v_j)$; it could be based on geodesic distance [7]. Formally, $simS(v_i, v_j) = \frac{1}{1+disS(v_i,v_j)}$. See Sect.4.2 below.

4.2. Semi Hybrid clustering

Semi-hybrid techniques combine simultaneously structural and attribute similarities through a weighted function as in Eq.1. **W-Cluster** and Combe's Model [3] are typical instances of this technique.

$$disG(v_i, v_j) = \alpha disT(v_i, v_j) + \beta disS(v_i, v_j) \tag{1}$$

$disT$ and $disS$ denote euclidean distance for attribute data and geodesic distance for structure data respectively. A straightforward way to integrate link semantic is to combine relational, attribute and semantic similarities by adding another factor to the Eq.1 as described below.

4.3. Proposed Hybrid Clustering Model

To avoid confusion to that semi-hybrid method (not taking into account link direction), we add semantic property based on edge directionality named $simR(v_i, v_j)$ [8] and we call *semantic clusters* the groups detected from a directed attributed graph partitioning hybrid model. The proposed approach combines simultaneously 3 information data through a Node Attribute and Edge Directionality Similarity ($NAEDS$) as defined in Eq.2. Then,

we have applied $NAEDS$ in Louvain's method to find answer of the following question: *Whether semantic communities be detected by dealing with direction of the edges?*

$$simG(v_i, v_j) = \alpha simT(v_i, v_j) + \beta simS(v_i, v_j) + \gamma simR(v_i, v_j) \quad (2)$$

The equation Eq.2 computes a global Similarity $simG(v_i, v_j)$ between two vertices v_i and v_j by the linear combination of 3 measures respectively corresponding to each type of information. $simT(v_i, v_j)$ is the attribute based similarity. It is an arithmetic average between discrete attribute based similarity $simADiscr(v_i, v_j)$ (determined by counting the number of attribute values nodes have in common) and textual attribute based similarity $simA(v_i, v_j) = \frac{1}{1+\sqrt{\Sigma_d(w_i^d - w_j^d)^2}}$ based on the euclidean distance. $simS(v_i, v_j)$ corresponds to the relational based similarity (see Sect.4.1).
And $simR(v_i, v_j) = \frac{|\Delta_{ij}|}{|\Delta_j|} * \frac{|\Gamma_j^{in} \cap \Gamma_i^{in}|}{|\Gamma_j^{in} \cup \Gamma_i^{in}| - \theta}$ as defined by Gamgne et al. [8], represents edge directionality based similarity which focuses on triad density and neighbourhood of vertices. Then the global similarity measure is used as pairwise similarity measure in the Louvain's method to partition the graph into clusters. The objective is to evaluate the scalability of the method based on this global similarity by extracting semantic clusters. α, β and γ are weighting factors that enable to give more importance to the structural, attribute or semantic similarity. $\gamma = 1 - \alpha - \beta$ and $\alpha, \beta, \gamma \neq 0$.

5. Experimental Study

In this section, we performed extensive experiments to evaluate the performance of the linear combination-based approach on real-world network datasets. All experiments were done on a $2.3GHz$ *Intel Pentium IV* PC with $6GB$ main memory, running Windows 8. Python and R package were used for implementations.

5.1. Experimental Datasets and evaluation measures

To our knowledge, there is no referenced benchmark with relational and attributes information handling link semantic (edge directionality). We construct a small ground truth dataset, a food web network, in order to compare each vertex to its real cluster. So, two datasets for experiments are used :

Food web : A typical illustration dataset as shown in Fig.1 is case of food web network where a vertex represents a species and edge the relationship between prey and predator.

Political Blogs Dataset: A directed network of hyperlinks between weblogs on US politics. This dataset contains $1,490$ weblogs with $19,090$ hyperlinks between these webblogs. Each blog in the dataset has an attribute describing its political leaning as either *liberal* of *conservative*.

We use two measures of Density and Normalized mutual information (NMI) to evaluate the quality of clusters generated by different methods.

5.2. Assumptions on food web illustration

Here we enumerate partitioning scenario and present expected results. We consider 5 subsets of vertices A, B, C, D, E describing species diet mode and by their reproduction mode, to be real semantic cluster of the hybrid clustering. The Table 1. shows the described illustration network according to each property :

Table 1: Number of species by nutrition sector and mode of reproduction

Diet Mode		Mode of reproduction	Number
A	Carnivorous	Viviparous	8
B	Carnivorous	Oviparous	3
C	Herbivorous	Viviparous	7
D	Herbivorous	Oviparous	4
E	Vegetables	Asexual or sexual	3
Total			25

– Semi attribute semantic (Textual) : 3 clusters in which species are grouped by their mode of reproduction. The ground truth partition is formally defined as $P_a = \{A \cup C, B \cup D, E\}$.

– Semi Relational-semantic (Neighbourhood) : 3 clusters in which species are grouped by their diet mode. The ground truth partition is formally defined as $P_r = \{A \cup B, C \cup D, E\}$.

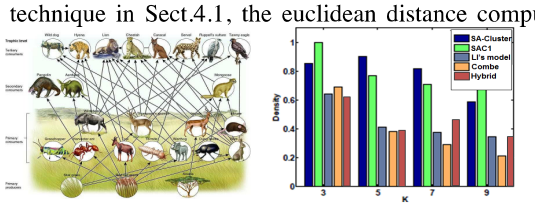
– Semantic : 5 clusters (species categories) : If we want to identify species by their both diet mode and mode of reproduction characteristics, then attributes(textual information), relational and directionality properties should be used. Like this, the resulting partition is $P_s = \{A, B, C, D, E\}$.

6. Model evaluations and results

6.1. Evaluation on illustration dataset

Given that this study focuses on directed attributed graphs which have not yet been investigated in detail, the evaluation consists in checking these assumptions described in Sect.5.2, by evaluating stated models of Sect.4 (M_a, M_r, SH_{ar}). We compare these 3 models (M) and (SH) with the hybrid model (H_s). The synthesis of results is shown in Table.2, according to the Normalized Mutual Information (NMI) measure [1]. Then clusters issued from the ground truth clustering transcripts the following partitions : the group of species by their diet mode (P_r), by their mode of reproduction (P_a), and by the both simultaneously (P_s).

– Clustering according to textual attributes : M_a Model. In this approach corresponding to the technique in Sect.4.1, the euclidean distance computed on the tex-



(a) Food web illustration network with Diet sectors

(b) Density comparison on Poltitical Blogs

Figure 1: Example of datasets and results

Table 2: Results : NMI

Models	P_r	P_a	P_s
M_r	**0.753**	0.350	0.323
M_a	0.741	**0.842**	0.625
SH_{ar}	$[0.028 - 0.291]$	$[0.205 - \mathbf{0.441}]$	$[0.085 - 0.397]$
H_s	$[0.098 - 0.217]$	$[0.110 - 0.185]$	$[0.558 - \mathbf{0.895}]$

tual attributes helps to weight each edge ; then an unsupervised method is applied to the resulting graph. The method performs well when the ground truth partition is $P_a = \{A \cup C, B \cup D, E\}$ by a higher NMI value (0.842) than considering the partitions P_r or P_s.

– Clustering according to relations : M_r Model. This method firstly exploits relations and secondly, with attributes handling, it detects communities so that the nodes in the same community are densely connected as well as homogeneous [2]. The NMI value for the ground truth partition $P_r = \{A \cup B, C \cup D, E\}$ is higher (0.753) than its value for the ground truth partition P_a and P_s. This result demonstrates that a technique based on successively relations then attributes, performs well in case of detecting two clusters of species with a densely internal connectivity, corresponding to diet mode.

– **Semi-hybrid attributed based clustering : SH_{ar} Model.** As far as this method is concerned, it deals with both types of information simultaneously as studied by Largeron [3] through a weighted distance function. In experiments, the NMI value fluctuates as a function of the weighting factors α and β. It changes its value according to the weighting factor α. NMI is in the interval $[0.028 - 0.291]$ for P_r ground truth and $[0.205 - 0.441]$ for P_a when α values are respectively 0.5 and 0.75. $\beta = 1 - \alpha$. SH_{ar} Model performs the best for the ground truth P_a, meaning that textual attributes describe better the vertices similarity, but produces weak outcomes as proved by [3] for the overall results.

– Hybrid attributed based clustering : H_s Model. The objective of this hybrid based experiment consists in 2 ways. First it shows that the consideration of the textual attributes improves better the cluster semantics through the highest NMI values as presented in bold in the Table.2. Second it shows that combining simultaneously the three types of information which are link semantic, relational and attribute properties respectively, leads to the highest NMI for that expected partition $P_s = \{A, B, C, D, E\}$. Like this, it detects the five classifying species clusters by their diet and reproduction mode simultaneously with a NMI value of 0.895 when the weighting factors α and β both equal 0.33; NMI value decreases to 0.558 when the weighting factors α and β equal 0.5 and 0.40 respectively, meaning that the negligence of the third factor relating to link semantic property affects the result.

6.2. Evaluation on Polblogs dataset

The Table 3 presents NMI for P_s partition, with $\alpha = \beta = \gamma = 0.33$, while the figure 1b compares Density for each model through the number of cluster. These results strengthen the interpretation according to that high density does not inevitably denote good separation of communities.

Table 3: Results : *Density*

Models	SAC1	SA-Cluster	Li's model	Combe's model	Hybrid model
NMI	0.153	0.350	0.323	0.675	**0.878**

7. Conclusion and future works

This work focused on the presentation of a hybrid clustering approach based on a proposed similarity. This measure takes into account 3 properties : semantic, relational and attributes. As presented below, we obtained different results according to the clustering technique and to the kind of data in the directed attributed food web graph we built.

An illustration on a food web network helped to underline the choice of each method relating to the kind of information (textual or numeric). The experiments show that on the one hand, the consideration of textual documents as attributes in the partitioning process leads to expected results based on the determination of species by their reproduction and nutrition modes simultaneously, and on the other hand, the properties strengthens the cluster semantic as computed through the NMI highest value. Nevertheless it has been difficult to integrate simultaneously two textual attributes relating to both reproduction mode and nutrition mode. For this reason, the second one has been processed as a numeric. Although this method is simple, it is hard to set/tune the parameters as well as interpret the weighted similarity function. Future works intend to apply large real-world networks and study weighting factors distribution.

8. References

[1] K. STEINHAEUSER, N. V. CHAWLA, "Identifying and evaluating community structure in complex networks", *Pattern Recognition Letters*, (2009).

[2] H. LI , Z. NIE , W. C. LEE, "Scalable Community Discovery on Textual Data with Relations", *ACM conference on Information and knowledge management*,pp. 1203-1212, (2008).

[3] D. COMBE, C. LARGERON , M. GERY, E. EGYED-ZSIGMOND "Détection de communautés dans des réseaux scientifiques à partir de données relationnelles et textuelles.", *MARAMI*, (2012).

[4] Y. ZHOU, H. CHENG , Y. JEFFREY XU "Graph Clustering Based on Structural/Attribute Similarities", *Adv. Intell. Data Anal.*, pp. 181-192 (2009).

[5] D. COMBE, C. LARGERON, M. GERY, E. EGYED-ZSIGMOND "I-louvain: An attributed graph clustering method.", *Adv. Intell. Data Anal. XIV*, pp. 181?192. Springer (2015).

[6] T. DANG, E. VIENNET "Community detection based on structural and attribute similarities.", *In: International Conference on Digital Society (ICDS)*, pp. 7-12 (2012).

[7] NEWMAN, M.E., GIRVAN M. " Detecting community structure in networks." *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38(2), pp. 321-330, 2004.

[8] F. GAMGNE , N. TSOPZE, R. NDOUNDAM, " Novel method to find directed community structures based on triads cardinality." *Proceedings of CARI'16.*, vol. 2016, pp. 8-15, (2016).

[9] F. GAMGNE , N. TSOPZE, " Communautés et rôles dans les réseaux sociaux." *Actes du CARI'14*, pp. 157-164, (2014).