Integrating heterogeneous information within a social network for detecting communities

Juan David Cruz Gomez, Cécile Bothorel, François Poulet

Integrating heterogeneous information within a social network for detecting communities

Juan David Cruz
Lab-STICC, UMR CNRS 3192
Dpt. LUSSI, Telecom-Bretagne
Technopôle Brest - Iroise, 29238 France
Tel: +33 (0)2 29 00 12 79
Email: juan.cruzgomez@telecom-bretagne.eu

Cécile Bothorel
Lab-STICC, UMR CNRS 3192
Dpt. LUSSI, Telecom-Bretagne
Technopôle Brest - Iroise, 29238 France
Tel: +33 (0)2 29 00 14 47
Email: cecile.bothorel@telecom-bretagne.eu

François Poulet
University of Rennes I – IRISA
Campus de Beaulieu Rennes, 35042 France
Tel: +33 (0)2 99 84 74 37
Email: francois.poulet@irisa.fr

***Abstract*—Attributed graphs can be described using two dimensions: first a *structural* dimension that contains the social graph, e.g. the actors and the relationships between them, and second a *compositional* dimension describing the actors, e.g. their profile, their textual publications, the metadata of the videos they share, etc. Each of these dimensions can be used to explain different phenomena occurring on the social network, whether from a connectivity or an thematic perspective. This paper claims that the integration of both dimensions would allow researchers to analyze real social networks from different perspectives. We present here a novel approach to the community detection problem with the integration of the two dimensions composing an attributed graph. We show how to integrate but also how to control the integration of two different partitions, one based on the links, the other one based on the attributes. The resulting partition exhibits interesting properties, such as dense and homogeneous groups of actors, revealing new types of communities to the analyst. Because we use a contingency matrix, and because the analyst may invent new ways of combining rows and columns, we open new perspectives for the exploration of attributed social networks.**

I. INTRODUCTION

According to Wasserman and Faust [1] social networks contain two different information dimensions that are represented by two variables: a structural one and a compositional one. The structural variable is used to describe the network in terms of the connections between actors, such as friendship. The composition variable describes each actor individually using their attributes such as origin, preferences, messages sent, topics of interest and/or other profile information. Such networks are also called attributed graphs. Each variable has been defined in different spaces making unfeasible their direct comparison. Additionally in the case of clustering, the approaches differ for each variable specifically regarding the quality measures, density for links, entropy for attributes.

We present a framework for detecting communities in attributed graphs. The objective is to detect communities of well connected *and* similar nodes. We propose a novel approach to integrate, but also control the integration of two different partitions, one based on the links, the other one based on the attributes using a contingency matrix. Such matrix describes the agreement between the partitions and by manipulating its rows and columns we can control the combination of partitions, and thus to explain social networks from both structural and compositional perspectives.

The paper is organized as follows: Section II presents relevant works in community detection for attributed graphs, in Section III the problem is introduced and some basic notation is presented, next in Section IV the algorithm is presented and Section V presents some experiments and discussion before the conclusion.

II. RELATED WORK

Several methods have been developed to detect communities in an attributed graph. Neville et al., [2] present a clustering approach that uses a similarity metric S_{ij} to change the weights of the edges of the graph and then find the communities. A similar approach has been presented by Steinhaeuser et al., [3]. The difference between these methods lies on the similarity function and the node selection process. Cruz et al., [4] present an entropy based algorithm while Zhou et al., [5] present a random walk based approach.

III. PROBLEM DEFINITION

Let $\mathcal{S}(G, F^*)$ be an attributed graph. The structural variable is represented as a graph $G(V, E)$, where V and E are the set of nodes and edges respectively. The composition variable is represented by a set F^* of attributes . Let $\mathbf{C}_G$ be a partition of the nodes according to G and let $\mathbf{C}_{F^*}$ be a partition of the nodes according to the attributes. We assume here that the partitions have been discovered by previous treatments (e.g. clustering), or they might have been given by declarative memberships (e.g. fan page).

Partitions $\mathbf{C}_G$ and $\mathbf{C}_{F^*}$ are expressed as affiliation matrices of size $|V| \times m$ and $|V| \times r$ respectively, where m is the number of structural groups and r is the number of compositional groups. The contingency matrix can be calculated as:

$$\mathcal{C} = \mathbf{C}_G^T \mathbf{C}_{F^*} \tag{1}$$

Each entry of the matrix $\mathcal{C}$ represents the number of common nodes between the structural group i and the compositional group j. To evaluate this work and compare partitions, we use the Adjusted Rand Index – ARI proposed by Hubert and Arabie [6], which provides a measure of the distance between two partitions.

IV. COMMUNITY DETECTION ALGORITHM

We have previously presented in [**?**] that using both structural and composition variables allows to consider the final partition as a refinement of one of the partitions in terms of the other, but we face a lack of control on the process and the results. We want here to generalize the approach and offer a control on the integration step. Our new community detection algorithm takes advantage of the configuration of the partitions through the matrix $\mathcal{C}$, which represents the relationships between our two partitions $\mathbf{C}_G$ and $\mathbf{C}_{F^*}$.

Algorithm 1 outlines the integration of two partitions via a row manipulation proposal. The algorithm starts by generating the contingency matrix (line 2). Then each row, corresponding to a

```
   Data: $\mathbf{C}_G$, $\mathbf{C}_{F^*}$
   Result: $\mathbf{C}^*$
1  $\mathcal{C}^* \leftarrow \emptyset$;
2  $\mathcal{C} \leftarrow \mathbf{C}_G^T \mathbf{C}_{F^*}$;
3  $i \leftarrow 0$;
4  while $i < rows\,(\mathcal{C})$ do
5      $\mathbf{C}_i \leftarrow$ row_process $(\mathcal{C}_{i\cdot})$;
6      $\mathcal{C}^* \leftarrow \mathcal{C}^* \oplus \mathbf{C}_i$;
7      $i \leftarrow i + 1$;
8  end
9  $\mathbf{C}^* \leftarrow$ rebuild_partition $(\mathcal{C}^*)$;
10 return $\mathbf{C}^*$
```

Algorithm 1: Row manipulation community detection algorithm

structural community i, is processed (line 5) in order to allocate the common nodes with the compositional communities j into one or several sub-communities. This process produces a matrix $\mathbf{C}_i$ of $s \times r$, where $1 \leq s \leq r$ is the number of subgroups that can be created from the $i-$th structural community. This matrix $\mathbf{C}_i$ is concatenated (line 6) to the matrix $\mathbf{C}^*$ that contains the new configuration of the final partition.

The division of each row (line 5) can be made according to different criteria. And it is how this method controls the integration of variables. Here we address intrinsic criteria, depending only on the configuration of the contingency matrix. In this paper we show two possible approaches: in a *naïve approach*, for each row $\mathcal{C}_i$, if $\mathcal{C}_{ij} > 0$ then the nodes belonging both to the structural group i and the composition group j will form a community; in the *variance-based approach*, for each element j of $\mathcal{C}_i$, if $\frac{(\mathcal{C}_{ij} - \mu_i)}{\sigma_i} \geq 1$ that element will be a new community. Here μ_i and σ_i are the average value and the standard deviation of the row i respectively.

V. Experiments and Results

To test the algorithm we have performed experiments on several attributed graphs. Here are the results for a personal Facebook dataset, containing 334 nodes and 5394 edges. The attributes are manually extracted from the explicit profiles (academic and professional skills). The dataset gathers people known at work, in university, family, etc.

Two partitions have been derived: first the structural partition $\mathbf{C}_G$, generated with the links, by the Louvain method [8] which is designed to optimize the modularity; second, the compositional partition $\mathbf{C}_{F^*}$ is the result of an unsupervised clustering on the attributes with Self-Organizing Maps – SOM [9] that optimizes a distance measure such as the Euclidean distance. $\mathbf{C}_G$ contains 6 groups (mainly related to different periods of life) while $\mathbf{C}_{F^*}$ contains 7 groups (skills oriented).

The ARI of our contingency matrix is 0.0189, which is low and confims orthogonal dimensions. We next apply our algorithm 1 with both the naïve and the variance integration methods and compare the integrated partition in terms of density (to compare with the pure structural partition $\mathbf{C}_G$) and entropy (to compare with the pure compositional partition $\mathbf{C}_{F^*}$).

With the naïve integration, each structural group i is divided into the number of composition groups with size greater than zero, in this case leading to 40 groups. As expected, since all the structural groups have been divided according to each attribute, we maximize the compositional quality (entropy is 0) while dropping the density value (decreasing social clustering).

Partition	**Groups**	**ARI** (w.r.t $\mathbf{C}_{F^*}$)	**Density**	**Entropy**
$\mathbf{C}_G$	6	0.0189	0.9718	15.1475
$\mathbf{C}^*_{\text{Naïve}}$	40	0.2819	0.1294	0
$\mathbf{C}^*_{\text{Variance}}$	12	0.1063	0.651093	4.5502

TABLE I. Summary of results of the algorithm for the Facebook social network

The variance integration method will extract only the more representative compositional communities within the structural groups. It reveals more interesting results, and shows the interest of controlling the integration of partitions. The division of the communities according to the skills shows some homogeneous groups and in general strong categories within the structure. For example the skill in Software Engineering is present in almost each group with an important number of members. This kind of information can be hidden within the structure of the graph when the composition information is discarded.

VI. Conclusion and future work

We present in this paper a novel approach to the community detection problem that integrates the two kinds of variables contained in an attributed social network. This approach takes advantage of the summarization of the two variables of the social network made with the contingency matrix. This matrix contains the agreements between two partitions issued from different types of information, making them comparable.

The rows of the contingency matrix represent the groups of the structural partition while the columns represent the groups of the compositional partition; therefore manipulating the rows in function of the columns yields to a new partition configuration that integrates information from the composition variable.

Once the contingency matrix has been found, the algorithm explores each row to determine whether it is possible to decompose it into several sub-communities. We proposed two ways to do this, first a naïve method that converts every non-zero entry of the contingency matrix into a new community: these communities are composed of nodes of one type only. Second a method based on the variance of each composition category composing the structural community: the new communities are created from separating from the original community those that contribute the most to the variance. This last criterium allows us to decompose the structural partition in terms of the composition variable while keeping a good tradeoff between the density and entropy.

The decomposition is a controlled process and we show here how an analyst could choose different criteria or strategies to combine the dimensions. This work is a very preliminary research, and future work includes new row division methods, but also how to select the structural groups to divide.

References

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. No. 8 in Structural Analysis in the Social Science, Cambridge University Press, 1994.

[2] J. Neville, M. Adler, and D. D. Jensen, "Clustering relational data using attribute and link information," in *Proceedings of the Workshop on Text Mining and Link Analysis, Eighteenth International Joint Conference on Artificial Intelligence*, (Acapulco, Mexico), 2003.

[3] K. Steinhaeuser and N. Chawla, "Community detection in a large real-world social network," in *Social Computing, Behavioral Modeling, and Prediction* (H. Liu, J. Salerno, and M. Young, eds.), pp. 168–175, Springer US, 2008.

[4] J. D. Cruz, C. Bothorel, and F. Poulet, “Entropy based community detection in augmented social networks,” in *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, pp. 163 –168, oct. 2011.

[5] Y. Zhou, H. Cheng, and J. X. Yu, “Graph clustering based on structural/attribute similarities,” *Proc. VLDB Endow.*, vol. 2, pp. 718–729, August 2009.

[6] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, pp. 193–218, 1985. 10.1007/BF01908075.

[7] J. D. Cruz, C. Bothorel, and F. Poulet, “Détection et visualisation des communautés dans les réseaux sociaux,” *Revue d’Intelligence Artificielle*, vol. 26, no. 4, pp. 369–392, 2012.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008 (12pp), 2008.

[9] T. Kohonen, *Self-Organizing Maps*. Springer, 1997.