Integrating heterogeneous information within a social network for detecting communities
Juan David Cruz Gomez, Cécile Bothorel, François Poulet

To cite this version:
Juan David Cruz Gomez, Cécile Bothorel, François Poulet. Integrating heterogeneous information within a social network for detecting communities. ASONAM 2013: the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Aug 2013, Niagara Falls, Canada. 2013. <hal-00857225>

HAL Id: hal-00857225
https://hal.archives-ouvertes.fr/hal-00857225
Submitted on 11 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Integrating heterogeneous information within a social network for detecting communities

Juan David Cruz
Lab-STICC, UMR CNRS 3192
Dpt. LUSI, Telecom-Bretagne
Technopolis 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. LUSI, Telecom-Bretagne
Technopolis Brest - Iroise, 29238 France
Tel: +33 (0)2 29 00 14 47
Email: cecile.bothorel@telecom-bretagne.eu

François Poulet
University of Rennes 1 – IRISA
Campus 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 a 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 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 $S(G, F^*)$ 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 $C_G$ be a partition of the nodes according to $G$ and let $C_{F^*}$ 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 $C_G$ and $C_{F^*}$ 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:

$$\mathbf{C} = \mathbf{C}_G \mathbf{C}_{F^*}$$

Each entry of the matrix $\mathbf{C}$ 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 [7] 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 $\mathbf{C}$, which represents the relationships between our two partitions $C_G$ and $C_{F^*}$.

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
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 naive method that converts every non-zero entry of the contingency matrix into a new community: these communities are created from separating from the original structural groups to divide.

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


