

Clustering of Links and Clustering of Nodes: Fusion of Knowledge in Social Networks

Erick Stattner and Martine Collard

Abstract The extraction of knowledge from social networks is an area that has experienced significant growth in recent years. Indeed, thanks to the improvement of storage and calculation capacities, and the heterogeneity of data that can currently be extracted, much effort has been made to go beyond traditional knowledge, by proposing new kinds of patterns that take into account the context. However, while many works were interested in designing new patterns of knowledge or in optimizing existing approaches, few studies have been focused in merging patterns and on the useful knowledge emerging from such fusions. In this work, we focus on two network clustering approaches, able to extract two distinct kinds of patterns, and we seek to understand both the intersections that can exist between them and the knowledge that emerges from their fusion. The first is the classical nodes clustering approach that consists in searching for communities into a network. The second is the search for frequent conceptual links, a new link clustering approach that aims identifying frequent links between groups of nodes sharing common attributes. We propose a set of original measures that aim to evaluate the amount of shared information between these patterns when they are extracted from a same network. These measures are applied to three datasets and demonstrate the interest in simultaneously considering several sources of knowledge.

1 Introduction

The domain of knowledge extraction from social networks, also called *social network mining* (Getoor and Diehl 2005; Scott 2011), has experienced strong growth in recent years. While pioneering works have proposed various methods to address classical data mining tasks such as classification of nodes, prediction of links or clustering of nodes, recent approaches have attempted to go beyond traditional knowledge

E. Stattner (✉) · M. Collard
LAMIA Lab., University of the French West Indies, Pointe-à-Pitre, France
e-mail: erick.stattner@univ-ag.fr

M. Collard
e-mail: martine.collard@univ-ag.fr

patterns by defining new kinds of knowledge suitable to the context (Manyika et al. 2011). Indeed, thanks to the improvement of storage and computation capacities, and the heterogeneity of data that can currently be extracted from online systems, more and more works have focused on approaches combining several sources of data, redefining traditional patterns of knowledge.

Clustering from social networks has been an active research area that has received a lot of contributions. Indeed, in natural or social systems, entities often tend to organize themselves in groups (Croft et al. 2008). For example, we observe that sharing common interests leads to the emergence of online communities through discussion forums or the exchange of messages or files. The detection of such groups is a good way to identify substructures that possibly have major roles in the targeted systems. Thus, the identification of these clusters and the comprehension of mechanisms underlying their formation are relevant challenges in many disciplines for uncovering relationships between the structure and the function into complex systems.

First clustering approaches exploited only the structure of the network in order to identify some particular patterns called *communities* (Radicchi et al. 2004; Fortunato 2009), namely groups of nodes densely connected. More recently, new approaches have attempted to combine both the network structure and the properties of nodes (Zhou et al. 2009; Stattner and Collard 2012b).

Nevertheless, the great majority of these works is conducted without taking into account the complementarity of the knowledge that can be acquired. Indeed, while many works were interested in designing new kinds of knowledge or in optimizing existing approaches, few studies have been focused in the fusion of patterns and the knowledge that could emerge from such fusions.

In this paper, we focus on two network clustering approaches and we seek to understand both the intersections that can exist between them and the useful knowledge emerging from their fusion. First, the classical *node clustering* approach that consists in searching for communities into a network. Second, the search for frequent conceptual links (FCL), a *link clustering* approach that exploits both the network structure and the properties of nodes to identify frequent links between groups of nodes sharing common attributes.

Our objective is to evaluate the potential relationships existing between FCL and communities for understanding how the patterns obtained with both approaches may overlap. For this purpose, we propose a set of original measures that aim to evaluate the amount of shared information between these patterns when they are extracted from the same network. These measures are then applied to three datasets (a proximity-based network, a product co-purchasing network and a phone call network) for demonstrating the interest to consider simultaneously several sources of knowledge. For each network, we provide several examples of the knowledge resulting from the fusion.

The paper is organized as follows. Section 2 presents the related works conducted on the identification of clusters. Section 3 describes the notions of communities and frequent conceptual links and discusses the questions raised when they are combined. Section 4 is devoted to the measures proposed to evaluate the quality of the fusion

between communities and FCL. Section 5 presents the experimental results we have obtained by applying the measures to three datasets. Finally, Sect. 6 concludes and presents future directions.

2 Related Works

Numerous methods for identifying clusters from networks can be found in the literature (Riadh et al. 2009; Steinhaeuser and Chawla 2010; Yang et al. 2013). While these methods are all able to highlight groups from data arising from networks, we observe that some factors such as the extracted knowledge or the data used vary from one method to another. Several criteria can be used to classify these approaches. In this section, we present the two clustering methods addressed in this paper: the identification of communities and the search for frequent conceptual links, according to three main criteria.

(i) Extracted knowledge. The identification of communities and the search for frequent conceptual links provide two distinct kinds of patterns. On the one hand, the concept of community is currently the most common approach for *clustering nodes* in networks. It provides an information on groups of nodes most densely connected in the network (Newman 2006). The associated algorithms aim to partition the network in several connected components, called “*communities*”, so that the nodes in each component have a high density of connection while nodes in different components have a lower link density (Fortunato 2009).

On the other hand, frequent conceptual links provide an information on groups of nodes most frequently connected in the network, in which each group is defined as a set of nodes sharing common attributes. Here, “*conceptual*” means that such a link is not a real social link, but represents a “*meta-link*” that is a set of social links between two groups of nodes considered as a concept according to the formal concept analysis area (Ganter et al. 2005). The set of frequent conceptual links extracted from a network provides a “*conceptual view*”, namely a new network structure in which a node represents a group of nodes sharing common attributes and a link represents a frequent connection between two groups in the original network.

(ii) Clustering criterion. In the domain of group identification, the building of clusters may rely on various clustering criteria (Mangiameli et al. 1996; Lancichinetti et al. 2008). In traditional network clustering, approaches attempt to identify a network partition in which the number of inter-clusters links is maximized while the number of intra-clusters links is minimized. For this purpose, they use the criterion of *modularity* introduced by Newman (2006) to evaluate the quality of the partition. The modularity measures the density of links into a group and is commonly used as an optimisation function in some network clustering algorithms (Lehmann and Hansen 2007; Blondel et al. 2008). Some approaches perform clustering on networks by using different measures, such as those using Potts models (Kumpula et al. 2007).

The search for frequent conceptual links relies on the notion of *support*, well known about frequent itemsets (Agrawal and Srikant 1994). It allows to evaluate the

percentage of links in the network connecting a group of nodes satisfying a given property A to another group of nodes satisfying a given property B . Thus, the higher the value of support is, the higher the amount of links connecting nodes satisfying A to nodes satisfying B is Stattner and Collard (2012b).

(iii) Source of data. In several applications, networks are modeled by links and nodes may have various kinds of associated attributes. Such networks are called “*information networks*” or “*networks with content*” (El Gamal and Kim 2011). For instance, in a telecommunication network, consumers (nodes) may be identified by attributes such as age, type of package, job status, etc. If the wide majority of network clustering approaches does not take into account the attributes of nodes, some recent works have proposed new definitions of community for including node properties in the clustering task (Yoon et al. 2011). These approaches aim to provide a semantic decomposition of the network by focusing on the “*densely connected groups of nodes with homogeneous attributes values*” as explained in Zhou et al. (2009).

The search for frequent conceptual links exploits both network structure and node attributes (Stattner and Collard 2012a). The extracting process involves two key steps: a clustering phase, that builds the concepts by grouping nodes with common attributes and an evaluation phase that exploits the network links to assess the frequency of links between concepts.

3 Towards a Fusion of Knowledge

This section describes formally the concepts of *communities* and *frequent conceptual links*. We first present each kind of pattern, then we discuss the useful knowledge resulting from their fusion.

First of all, let $G = (V, E)$ be a social network, where V is the set of nodes (vertices) and $E \subseteq V \times V$ the set of social links (edges).

3.1 Communities in Social Networks

We define C as the set of communities extracted from the network G . We assume that there are no overlapping communities, thus a node belongs to one and only one community. We denote $F : V \rightarrow C$, the function that returns, for a given node v , the community to which it belongs.

The communities are extracted in order to maximize the modularity Q defined as follows:

$$Q = \frac{1}{2|E|} \sum_{ij} [W_{ij} - \frac{k_{v_i} k_{v_j}}{2|E|}] \delta(F(v_i), F(v_j))$$

where W_{ij} represents the weight of the edge between nodes v_i and v_j , k_{v_i} corresponds to the degree of node v_i and the δ -function is equal to 1 if $F(v_i) = F(v_j)$ and 0 otherwise.

The method we use in our experiments is the algorithm proposed by Blondel et al. (2008), based on modularity optimization.

3.2 Frequent Conceptual Links in Social Networks

V is defined as a relation $R(A_1, \dots, A_p)$ where each A_i is an attribute. Thus, each vertex $v \in V$ is defined by a tuple (a_1, \dots, a_p) where $\forall q \in [1 \dots p]$, $v[A_q] = a_q$, the value of the attribute A_q in v and $|R| = p$.

An item is a logical expression $A = x$ where A is an attribute and x a value. The empty item is denoted \emptyset . An itemset is a conjunction of items for instance $A_1 = x$ and $A_2 = y$ and $A_3 = z$. An itemset which is a conjunction of k non empty items is called a k -itemsets.

Let m and sm be two itemsets. If $sm \subseteq m$, we say that sm is a sub-itemset of m and m is a super-itemset of sm . For instance $sm = xy$ is a sub-itemset of $m = xyz$. Any itemset is a sub-itemset of itself.

We denote I_V the set of all itemsets built from V .

Let us consider G as a *unipartite directed graph*. Thus, for any itemset m in I_V , we denote V_m the set of nodes in V that satisfy m and we define:

- the *m-left-hand linkset* LE_m as the set of links in E that start from nodes satisfying m i.e.
 $LE_m = \{e \in E ; e = (a, b) \quad a \in V_m\}$
- the *m-right-hand linkset* RE_m as the set of links in E that arrive to nodes in V_m i.e.
 $RE_m = \{e \in E ; e = (a, b) \quad b \in V_m\}$

Definition 1 (*Conceptual link*) For any two elements m_1 and m_2 in I_V , the *conceptual link* (m_1, m_2) of G is the set of links connecting nodes in V_{m_1} to nodes in V_{m_2} .

For instance, if m_1 is the itemset cd and m_2 is the itemset efj , the *conceptual link* $(m_1, m_2) = (cd, efj)$ includes all links in E between nodes in V that satisfy the property cd with nodes in V that satisfy the property efj .

Let L_V be the set of conceptual links of $G = (V, E)$ and (m_1, m_2) be any element in L_V .

$$\begin{aligned} (m_1, m_2) &= LE_{m_1} \cap RE_{m_2} \\ &= \{e \in E ; e = (a, b) \quad a \in V_{m_1} \text{ and } b \in V_{m_2}\} \end{aligned}$$

Definition 2 (*Support of conceptual link*) We call *support* of any element $l = (m_1, m_2)$ in L_V , the proportion of links in E that belong to l .

$$\text{supp}(l) = \frac{|(m_1, m_2)|}{|E|}$$

For an itemset m and a conceptual link l , if $l = (\emptyset, m)$ or $l = (m, \emptyset)$ then $\text{supp}(l) = 0$.

Definition 3 (*Frequent Conceptual Link*) Given a real number $\beta \in [0 \dots 1]$, a conceptual link l in L_V is *frequent* if its support is greater than a minimum link support threshold β ,

$$\text{supp}(l) > \beta$$

Let FL_V be the set of frequent conceptual links (FCL) in $G = (V, E)$ according to a given link support threshold β .

$$FL_V = \bigcup_{m_1 \in I_V, m_2 \in I_V} \{(m_1, m_2) \in L_V ; \frac{|(m_1, m_2)|}{|E|} > \beta\}$$

Definition 4 (*Conceptual sub-link*) Let two any itemsets sm_1 and sm_2 be respectively sub-itemsets of m_1 and m_2 in I_V . The conceptual link (sm_1, sm_2) is called conceptual *sub-link* of (m_1, m_2) .

Similarly, (m_1, m_2) is called conceptual *super-link* of (sm_1, sm_2) .

We write $(sm_1, sm_2) \subseteq (m_1, m_2)$

Definition 5 (*Maximal frequent conceptual link*) Let β be a given link support threshold, we call *maximal frequent conceptual link* (MFCL), any frequent conceptual link l such as, there exists no super-link l' of l that is also frequent.

More formally, $\nexists l' \in FL_V$ such as $l \subset l'$.

MFCLs provide a conceptual view of the social network about groups of nodes that share common *internal* properties (or concepts according to the area of formal concept analysis (Ganter et al. 2005)) and that are the most connected. More precisely, the conceptual view is a graph structure in which each node is related to an itemset (i.e. group of nodes that satisfy this itemset), and each link corresponds to a MFCL. By this way, the conceptual view provides a semantic and reduced representation of the initial network. More precisely, the set of the maximal frequent conceptual links provides a conceptual and synthetic view of the social network in which only relevant links between groups of nodes are represented.

Definition 6 (*Conceptual view of the social network*) Let $G = (V, E)$ be a social network and β the minimum support threshold. We define G_β^* , the graph (M, L) , as the conceptual view of the network G obtained with the link support threshold β .

- M is the set of itemsets, called “*meta-nodes*”
- L is the set of maximal frequent conceptual links.

3.3 Merging Communities and Frequent Conceptual Links

Figure 1 shows resulting patterns extracted by community extraction and search for frequent conceptual links methods from a reference network. We can observe that patterns extracted by both methods provide two very different kinds of knowledge. The identification of communities extracts **cluster of nodes** based on the density of internal links, while the search for frequent conceptual links extracts **clusters of links** based on their frequency in the network. Obviously considering simultaneously these two kinds of pattern can improve the knowledge of these structures. It also raises a variety of interesting questions on the organisation of the involved structures such as:

1. Are communities composed by a single meta-node, i.e. a unique property?
2. Do the meta-nodes contain nodes that belong to a same community?
3. Do the frequent conceptual links connect nodes belonging to a same community, or nodes belonging to different communities?

To answer these questions related to the fusion between both kinds of pattern, we present in the next section a set of interestingness measures designed to evaluate the quality of the merging. More precisely, the proposed measures evaluate the degree of inclusion of communities in meta-nodes, and inversely, the degree of inclusion of meta-nodes in communities.

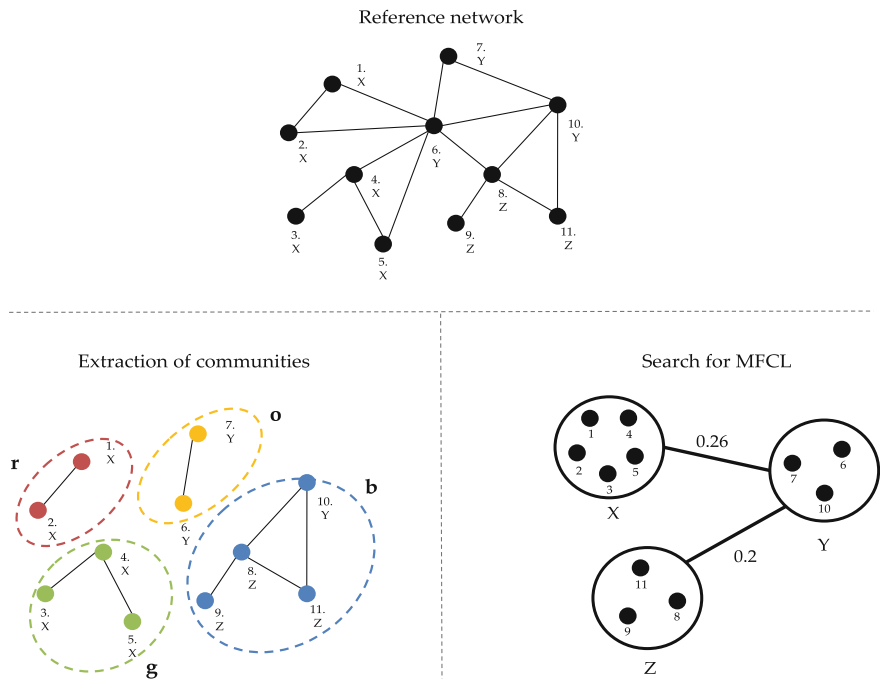


Fig. 1 Communities and maximal frequent conceptual links extracted from a reference network

4 Interestingness Measures

This section is devoted to the measures we propose for evaluating the intersections of both patterns: communities and frequent conceptual links.

4.1 Preliminaries

We remind that $G = (V, E)$ is a social network in which, V is the set of nodes and E the set of links. Cardinality of the sets V and E , respectively denoted $|V|$ and $|E|$ provides the number of nodes and the number of links.

C is the set of communities identified on the network by using a classical link-based clustering techniques (Blondel et al. 2008). Cardinality $|C|$ provides the total number of communities identified on the network.

We note V_c the set of nodes in V that belong to the community c , i.e. $V_c = \{v \in V ; F(v) = c\}$.

Finally, let $G_\beta^* = (M, L)$ be the conceptual view obtained by extracting maximal frequent conceptual links from the network G . The set M is the set of meta-node and $L \subseteq M \times M$ is the set of maximal frequent conceptual links. The extraction of MFCL from G can be performed by algorithms proposed in (Stattner and Collard 2012b). Let us specify that the computation time related to the extraction of frequent conceptual links exponentially increase with the number of links in the network. However, some works have been carried out in order to reduce the computation time by using some properties of node sets (Stattner and Collard 2013).

Let $m \in M$ be a given itemset. We remind that V_m is the set of nodes in V satisfying the property m .

In this paper, our objective is to understand the possible relationships between the patterns extracted with methods focusing on both communities and conceptual links. In a more semantic way, we investigate the relationships between densely connected groups of nodes (i.e. communities or clusters) and groups of nodes sharing common properties that are frequently connected in the whole network (i.e. frequent conceptual links).

For this purpose, three objects have to be considered: (i) *communities*, that are related to link-based clustering techniques and (ii) *meta-nodes* and (iii) *frequent conceptual links*, that refer to the patterns extracted by the conceptual links extraction techniques.

In this section, we present various measures, related to the homogeneity into each kind of objects to understand how communities are included in conceptual links, and inversely, how conceptual links are involved into communities.

4.2 Homogeneity Rate into a Community

The *homogeneity rate into a community*, noted H_c , is a measure that indicates, for a given community $c \in C$, its ability to aggregate nodes that belong to the same meta-node, i.e. a set of nodes sharing common properties. This measure corresponds to the fraction of meta-nodes that do not occur in the community c .

$$H_c = 1 - \frac{|\{m \in M ; \exists v \in V \text{ with } F(v) = c \text{ and } v \in V_m\}|}{|M|} \tag{1}$$

If $H_c = 0$, all meta-nodes are present in community c . More semantically, nodes in the community c satisfy all properties involved in conceptual links. Inversely, a high H_c value indicates that nodes in community c only belong to a small fraction of meta-nodes, i.e. nodes in community c tend to have similar properties.

For instance, the homogeneity rate in community r is $H_r = 0.6$, while the homogeneity rate in community b is $H_b = 0.3$ (see Fig. 2).

For considering weighting of a property into a community, we introduce $H_{c/m}$, the *homogeneity rate of a given meta-node m into a community c* . It corresponds to the fraction of nodes satisfying property m in community c .

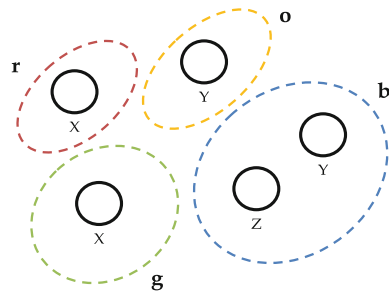
$$H_{c/m} = \frac{|\{v \in V ; F(v) = c \text{ and } v \in V_m\}|}{|\{v \in V ; F(v) = c\}|} \tag{2}$$

Thus if $H_{c/m} = 0$, nodes in meta-node m are not present in c . In a more semantic view, the nodes satisfying property m do not belong to community c . Inversely, when $H_{c/m}$ tends to 1, property m is satisfied by a high percentage of nodes in community c .

For instance, the homogeneity rate of meta-node X in community r is $H_{r/X} = 1$ (see Fig. 2). In the same way, the homogeneity rate of meta-node Z in community b is $H_{b/Z} = 0.75$.

As previously, the set of all $H_{c/m}$ values obtained for each pair (c, m) provides a $|C| \times |M|$ matrix.

Fig. 2 Meta-nodes (x, y and z) included into communities (r, g, o and b) from the example of Fig. 1



4.3 Homogeneity Rate into a Meta-node

The *homogeneity rate into a meta-node*, H_m , is a measure that indicates, for a given meta-node $m \in M$, its ability to aggregate nodes of the same community. It corresponds to the fraction of communities that do not occur in the meta-node m .

$$H_m = 1 - \frac{|\{c \in C ; \exists v \in V_m \text{ with } F(v) = c\}|}{|C|} \tag{3}$$

Thus, if $H_m = 0$, all communities are represented in the meta-node m . In other words, all communities contain nodes satisfying property m . Inversely, when H_m tends to 1, only a small percentage of communities is present in m , i.e. the meta-node contains nodes of the same community.

For instance, regarding the example of Fig. 1 containing 4 communities, the homogeneity rate into meta-node X (see Fig. 3) is $H_X = 0.5$, while homogeneity rate into meta-node Z is $H_Z = 0.75$.

To take into account the weighting, we introduce $H_{m/c}$, the *homogeneity rate of a given community c into a meta-node m* . This measure indicates the fraction of nodes of community c , in the meta-node m .

$$H_{m/c} = \frac{|\{v \in V_m ; F(v) = c\}|}{|V_m|} \tag{4}$$

Thus, if $H_{m/c} = 0$, nodes of community c are not present in m . More semantically, nodes in cluster c does not satisfy the property m . Inversely, when $H_{m/c}$ tends to 1, m is mostly represented in community c .

For example, starting from the example of Fig. 1, the homogeneity rate of community r in meta-node X is $H_{X/r} = 0.4$ (see Fig. 3). Similarly, homogeneity rate of community b in meta-node Z is $H_{Z/b} = 1$.

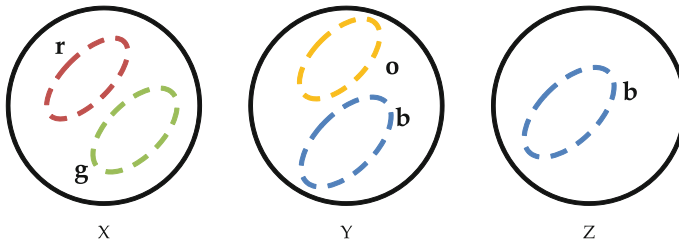


Fig. 3 Communities (r , g , o and b) included into meta-nodes (x , y and z) from the example of Fig. 1

4.4 Homogeneity Rate into a Conceptual Link

The *homogeneity rate* H_l , into a conceptual link, measures for a given frequent conceptual links $l = (m_1, m_2)$, its ability to connect nodes belonging to the same community. In other words, it indicates if nodes of the same community maintain a frequent conceptual link. It corresponds to the fraction of similar communities represented in both sides of the frequent conceptual links.

$$T_1 = \{c \in C ; \exists v \in V \text{ with } F(v) = c \text{ and } v \in V_{m_1}\}$$

$$T_2 = \{c \in C ; \exists v \in V \text{ with } F(v) = c \text{ and } v \in V_{m_2}\}$$

$$HL_l = \frac{|(T_1 \cap T_2)|}{|(T_1 \cup T_2)|} \tag{5}$$

Thus, for a given frequent conceptual link $l = (m_1, m_2)$, a low HL_l value indicates that nodes involved in both sides of the frequent conceptual link belong to different communities, while a high HL value indicates that a large amount of communities represented in meta-node m_1 are also represented in meta-node m_2 .

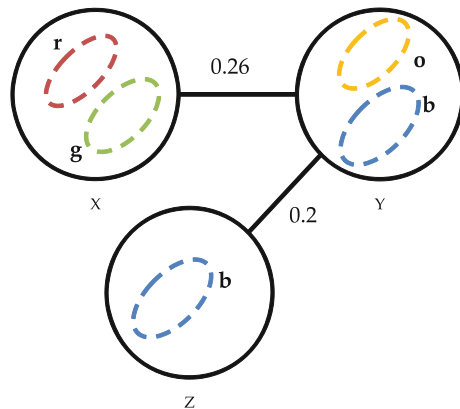
For example, the homogeneity rate into the conceptual link (Z, Y) is $H_{(Z,Y)} = 0.5$ (see Fig. 4). In the same way, the homogeneity rate into the conceptual link (X, Y) is $H_{(X,Y)} = 0$.

As previously, we introduce $H_{l/c}$, the *homogeneity rate of a given community c into the frequent conceptual link $l = (m_1, m_2)$* . More precisely, $H_{l/c}$ measures the difference in representation of a community c in meta-nodes m_1 and m_2 .

$$H_{l/c} = 1 - \frac{|H_{m_1/c} - H_{m_2/c}|}{\max(H_{m_1/c}, H_{m_2/c})} \tag{6}$$

Thus, for a given frequent conceptual link $l = (m_1, m_2)$, the homogeneity rate $H_{l/c} = 1$ indicates that the fraction of nodes of community c in meta-nodes m_1 and m_2 of l

Fig. 4 Communities (r, g, o and b) included into frequent conceptual links from the example of Fig. 1



is similar. Inversely, $H_{l/c} = 0$ indicates that at least one of the meta-nodes does not contain nodes belonging to the community c .

For example, the homogeneity rate of community b into the frequent conceptual link (Z, Y) is $H_{(Z,Y)/b} = 1 - \frac{0.7}{1} = 0.3$.

5 Experimental Results

We have conducted various set of experiments to evaluate the quality of the fusion. The results obtained show that very homogeneous structures can be found, and demonstrate the interest to consider simultaneously several sources of knowledge and more particularly the communities and the frequent conceptual links.

Section 5.1 describes the datasets used and their main characteristics regarding the network structural properties, the communities and their size and the frequent conceptual links and their properties. Section 5.2 presents and discusses the results we have obtained by applying the interestingness measures proposed on the three datasets.

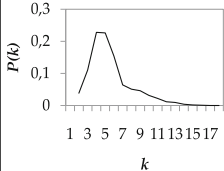
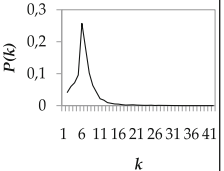
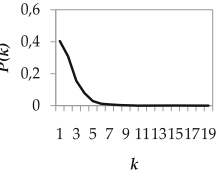
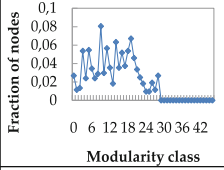
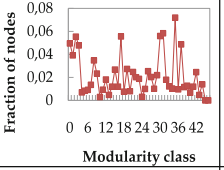
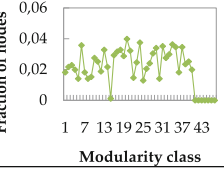
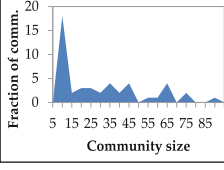
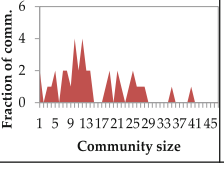
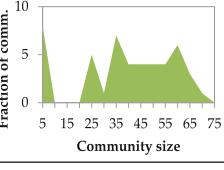
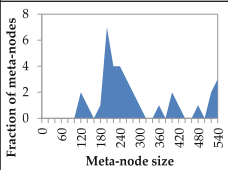
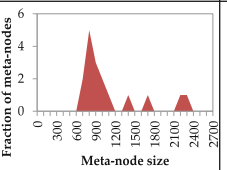
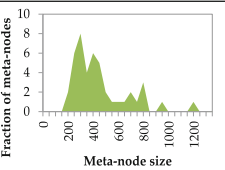
5.1 Testbed

Three datasets have been used in our experiments.

- (i) The first (referred as EpiSims in the remaining of the paper) is a geographical proximity-based social network obtained with *EpiSims* (Barrett et al. 2008), a simulation tool that statistically reproduces the daily movements of individuals in the city of Portland. In this network, two individuals are connected when they were co-located in the same place during the simulation.
- (ii) The second (referred as Amazon in the remaining of the paper) is a product co-purchasing network (Leskovec et al. 2007), extracted from the *Amazon* database, in which two products are connected when they were purchased together by a same user.
- (iii) The third (referred as Communications in the remaining of the paper) is a connected subnetwork of a very large communication network provided by a local mobile telephony operator (Stattner 2014) in French West Indies and Guiana. In this network, two individuals are connected when a telephone call was made between them.

The main characteristics of these datasets and the properties of the extracted patterns (communities and frequent conceptual links) are described in Table 1. The identification of the communities has been performed with the *Louvain Algorithm* proposed by Blondel et al. (2008), a nodes clustering method that relies only on the structure on the network. As the Louvain Algorithm is non-deterministic we

Table 1 Main properties of the dataset used

		EpiSims	Amazon	Communications	
General information	#Nodes	1043	5001	1705	
	#Attributes	6	7	7	
	#Links	2382	14981	1807	
	Density	0,0044	0,0012	0,0012	
	Coeff. Clust	0,7091	0,4874	0,0439	
	#Composante	1	1	1	
	Max degree	15	64	12	
	Avg degree	4,5675	5,9912	2,1196	
	Degree Distribution				
Communities	Modularity	0,864	0,883	0,945	
	#Clusters	29	45	40	
		Size of communities			
		Community size distribution			
Conceptual Links	#Meta-nodes	35	21	44	
	#FCL	116	43	105	
		Distribution size of meta-nodes			

focused on an extraction that represented a meaningful snapshot of communities. The extraction of maximal frequent conceptual links has been performed with the *MFCL-Min Algorithm* proposed in (Stattner and Collard 2012b). The minimum link support threshold β was set at 0.1, namely we keep only groups that contain at least 10 % of the network links.

(i) The EpiSims network is composed of 1043 nodes and 2382 links. Each node is identified by 6 attributes: (1) age class, i.e. $\lfloor \frac{age}{10} \rfloor$ (2) gender (1-male, 2-female), (3) worker (1-has a job, 2-has no job), (4) relationship to the head of household (1-spouse, partner, or head of household, 2-child, 3-adult relative, 4-other), (5) contact class (i.e. $\lfloor \frac{degree}{2} \rfloor$) and (6) sociability (i.e. 1-coeff. clust. > 0.5, 2-else). The network contains 29 communities, and 35 Meta-nodes and 116 frequent conceptual links have been identified.

Figure 5 shows the knowledge extracted from the Episims network: (a) Communities and (b) Conceptual view by keeping only FCL with $\beta \geq 0.2$ for more readability. In this figure, nodes that belong to a same community have an identical color. Moreover for simplicity, meta-nodes (properties) are denoted as follows:

$$(<att 1>, <att 2>, \dots, <att n>)$$

where $<att i>$ corresponds to the value of the attribute i on the node. The character ‘*’ means that the attribute may have any value.

For example, we can observe that the FCL $((*; 2; 2; *; *; *), (*; *; 2; *; *; *))$ has been identified with a support equals to 0.2. It indicates that 20 % of the links of the network connect women who has no job to individuals who has no job.

(ii) The Amazon network is composed of 5001 nodes and 14981 links. Each node is identified by 7 attributes: (1) product group (eg. Book, DVD, Video or Music), (2) number of similar co-purchased products (integer), (3) category (integer), (4) main category (eg. Literature and Fiction, Arts and Photography, Sport, ...), (5) sub cat-

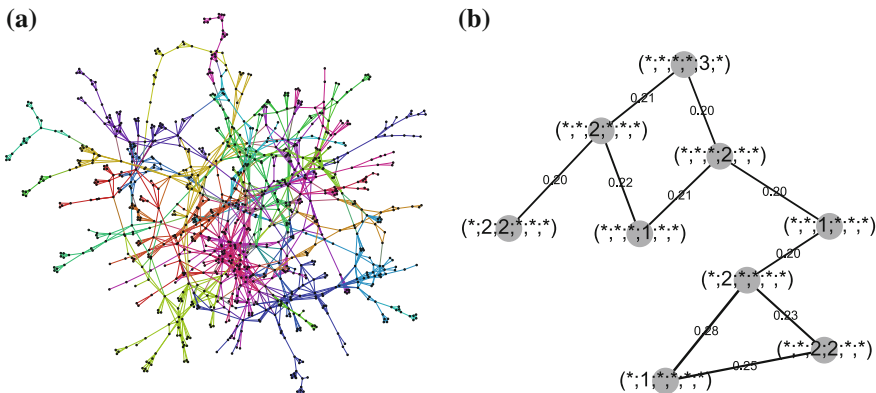


Fig. 5 Knowledge from Episims network: **a** communities and **b** conceptual view with $\beta \geq 0.2$

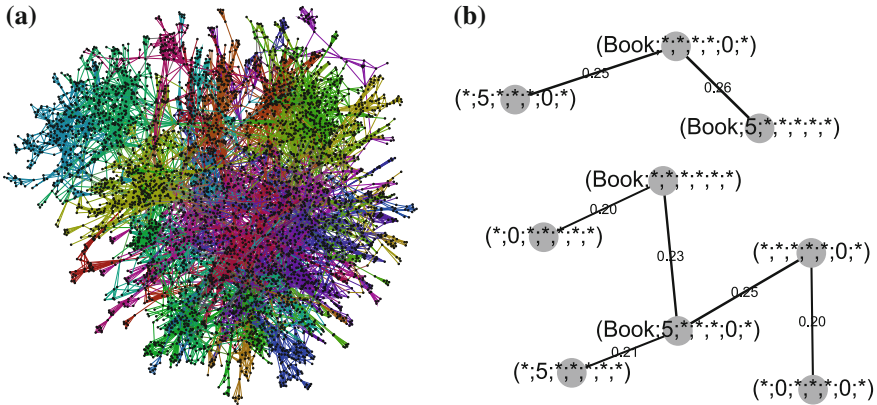


Fig. 6 Knowledge from Amazon network: **a** communities and **b** conceptual view with $\beta = 0.1$

egory (like (5)), (6) number of reviews (integer) and (7) rating (integer between 1 and 5). The network contains 45 communities and 21 Meta-nodes, and 43 frequent conceptual links have been identified.

Figure 6 shows the knowledge extracted from the Amazon network: (a) Communities and (b) Conceptual view by keeping FCL links with $\beta \geq 0.2$ for more readability. We can observe that the FCL ((Book; *, *, *, *, *, 0; *), (Book; 5; *, *, *, *, *, *)) is identified with a support of 0.26. It indicates that 26 % of the links of the Amazon network connect books that have no-review to books that are co-purchased with five similar products.

(iii) The Communication network is composed of 1705 nodes and 1807 links. The data have been processed to keep only calls between users, namely removing calls to voice mail, customer service, etc. Each node of this network is characterized by 7 attributes.

1. localisation (“Martinique”, “Guadeloupe”, “Guyane” or “Other”),
2. class of received calls number, i.e. $\lfloor \frac{\#received\ calls}{10} \rfloor$,
3. class of received average calls duration, i.e. $\lfloor \frac{rec.\ avg\ call\ duration}{10} \rfloor$,
4. class of outgoing calls number, i.e. $\lfloor \frac{\#outgoing\ calls}{10} \rfloor$,
5. class of outgoing calls average duration, i.e. $\lfloor \frac{out.\ avg\ calls\ duration}{10} \rfloor$,
6. class of number of SMS sent, i.e. $\lfloor \frac{\#SMS\ sent}{10} \rfloor$
7. class of number of SMS received, i.e. $\lfloor \frac{\#SMS\ received}{10} \rfloor$.

The network contains 40 communities and 44 Meta-nodes, and 105 frequent conceptual links have been identified.

Figure 7 shows the knowledge extracted from the Communication network: (a) Communities and (b) Conceptual view by keeping FCL links with $\beta \geq 0.2$ for more readability. We can observe that the FCL

$$(GUAD.; *, *, *, *, *, *) , (GUAD.; *, *, 1; *, *, *, *)$$

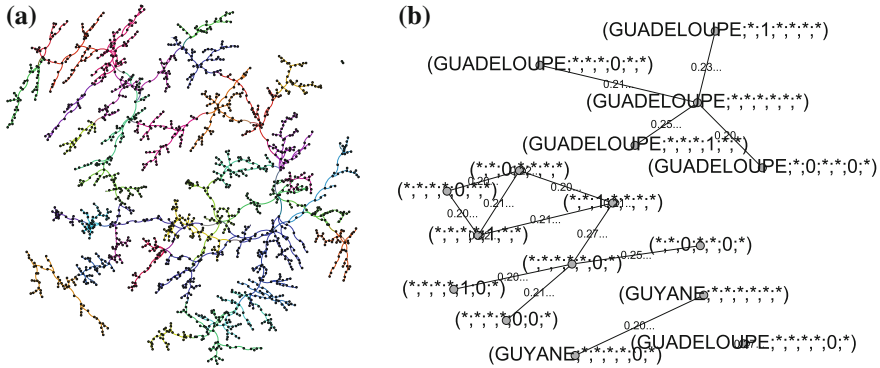


Fig. 7 Knowledge from Communication network: **a** communities and **b** conceptual view with $\beta = 0.1$

is identified with a support equals to 0.23. It indicates that 23 % of the links of the network connect consumers located in Guadeloupe to consumers located in Guadeloupe and having an average call duration comprised between 10 and 19 min.

Note that the datasets used are relatively small because of the difficulty for extracting FCL on large datasets. More particularly, in Stattner and Collard (2012b) it has been shown that the computation time exponentially increases with the number of attributes. However, some recent works have focused on the optimisation of the extraction process and have proposed various solutions to reduce the search space (Stattner and Collard 2013).

5.2 Results

In our experiments, we apply the proposed measures to the three datasets with the goal to identify homogeneous structures regarding the fusion of communities and frequent conceptual links. For this purpose we focus, for each measure, to the distribution of the values in order to highlight the amount of situations in which the measures are maximized. Moreover, for each measure we give some examples of interesting fusion.

5.2.1 Meta-Nodes Inside Communities

As a first step, Fig. 8 shows, for each dataset, the distribution of the homogeneity rate $H_{c/m}$ of a meta-node into a community. We remind that the homogeneity rate into a community allows evaluating if a community consists only of nodes belonging to the same meta-nodes, i.e. nodes satisfying common properties and involved in frequent conceptual links as described previously in Fig. 2.

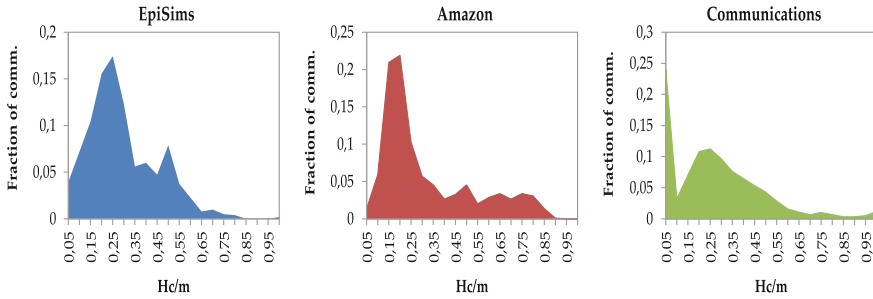


Fig. 8 Distribution of homogeneity rate $H_{c/m}$ of a meta-node into a community

We can observe that trends are very similar for the three networks. Indeed, for each dataset the vast majority of the homogeneity rates is rather low. For instance in the Episim network 91.13 % of the $H_{c/m}$ values are less than 0.5. In the Amazon network 81.16 % of the $H_{c/m}$ values are less than 0.5 and in the Communications network this proportion is 89.71 %. This result suggests that a strong proportion of communities are very heterogeneous in their structure, since they are not composed of nodes that belong to a same meta-node. Consequently, several attributes can be found in such communities.

However, our approach also allows highlighting that it exists a small percentage of communities which have a high homogeneity rate. For instance, in the EpiSims network 1.08 % of the $H_{c/m}$ values are higher than 0.75. These proportions are 7.93 % for the Amazon network and 4.09 % for the Communications network. This result indicates that it exists some communities very homogeneous since they are mainly composed of nodes belonging to a same meta-node, i.e. a group of nodes that share common attributes and that is involved in a frequent conceptual link.

Table 2 shows some interesting patterns regarding the $H_{c/m}$ measure. For example, 80 % of the nodes in the community 24 of the EpiSims network (see line 1 Table 2) is

Table 2 Examples of interesting patterns regarding $H_{c/m}$

Network	Community	Meta-Node	$H_{c/m}$
EpiSims	24 (10 nodes)	(*;*;2;*;3;*)	0.80
	13 (19 nodes)	(*;*;1;*;*)	0.73
	4 (25 nodes)	(*;1;*;*)	0.70
Amazon	39 (64 nodes)	(*;*;*;*;0;*)	0.87
	25 (50 nodes)	(Book;*;*;*;*)	0.84
	20 (46 nodes)	(*;5;*;*)	0.72
Communication	6 (31 nodes)	(GUADELOUPE;*;*;*;*)	1.00
	36 (59 nodes)	(GUYANE;*;*;*;*)	1.00
	29 (24 nodes)	(GUADELOUPE;*;*;*;0;*)	0.89

composed of nodes that belong to the meta-node (*; *; 2; *; 3; *), namely a group of individuals who have no job and have between 5 and 6 connections. In the same way, the community 29 of the Communication network is composed to 89 % of individuals located in the Guadeloupe island and sending between 0 and 9 SMS (see last line Table 2).

5.2.2 Communities Inside Meta-Nodes

In a second study, we have focused on the homogeneity rate $H_{m/c}$ of a community into a meta-node. As previously, we show on Fig. 9 the distribution of this measure for each dataset. We remind that the homogeneity rate into a meta-node allows evaluating if a meta-node (i.e. a group of nodes that share common properties and that is involved in a frequent conceptual link) is solely composed of nodes that belong to a same community. In other words, this measure assesses whether the nodes that share common attributes are densely interconnected.

The values obtained here for the homogeneity rate $H_{m/c}$ are very low whatever is the dataset. For instance, for the EpiSims network $max(H_{m/c})$ is 0.11, while it is 0.08 for the Amazon network and 0.11 for the Communication network. This result suggests that situations in which meta-nodes are fully homogeneous are very rare. In other words, it seems to be unlikely that the nodes into a meta-node are densely interconnected. Obviously, we can assume that these results vary according to the nature of the network and the semantics of the links.

Table 3 shows some examples of patterns regarding the $H_{m/c}$ measure. For instance, the first line of the table indicates that in EpiSims network 11 % of nodes satisfying property (1; 1; 2; 2; *; *) belong to community 20. In other words little boys who are between 0 and 9 years old are involved in frequent conceptual links and they are densely connected. In the same way, 11 % of the set of subscribers located in French Guiana whose received calls have a duration between 0 and 9 min and who have sent between 0 and 9 SMS (i.e. (GUYANE; *; 0; *; *; 0; *)) belong to community 28 (see line 7 Table 3).

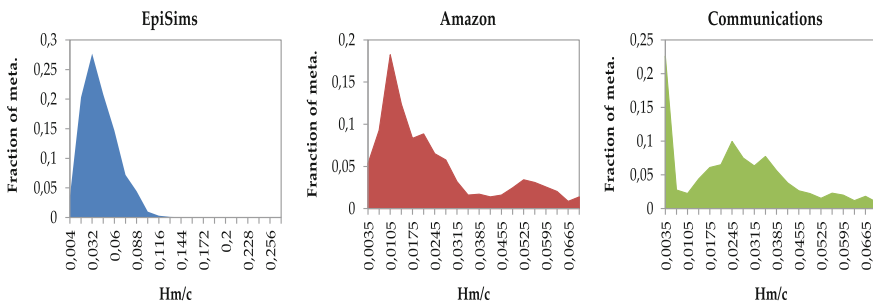


Fig. 9 Distribution of homogeneity rate $H_{m/c}$ of a community into a meta-node

Table 3 Examples of interesting patterns regarding $H_{m/c}$

Network	Meta-node	Community	$H_{m/c}$
EpiSims	(1;1;2;2;*,*) (119 nodes)	20	0.11
	(1;*,2;2;*,*) (194 nodes)	19	0.10
	(2;*,2;2;*,*) (196 nodes)	9	0.10
Amazon	(Book;0;*,*,*,0;*) (996 nodes)	35	0.08
	(*,*,*,*,General;*,*) (1071 nodes)	35	0.07
	(Music;*,*,*,*,*,*) (997 nodes)	35	0.07
Communication	(GUYANE;*,*0;*,*0;*) (212 nodes)	28	0.11
	(GUYANE;*,*,*,1;*,*) (300 nodes)	33	0.10
	(GUADELOUPE;*,*1;*,*1;*,*) (192 nodes)	10	0.09

5.2.3 Communities Inside Frequent Conceptual Links

In the last study, we have focused on the homogeneity rate $H_{l/c}$ of a community into a frequent conceptual link. Figure 10 shows the distribution of this measure for each dataset. We remind that the homogeneity rate into a frequent conceptual link measures the ability for a FCL to connect nodes that belong to same communities. It allows evaluating if a frequent conceptual link is composed, for right and left sides, to nodes belonging to a same community as described in Fig. 4.

We can observe that the trends are very similar for the three networks. Indeed, for each dataset the vast majority of the homogeneity rates obtained is rather high. For instance, in the EpiSims network 87.97 % of the $H_{l/c}$ values are greater than 0.75. In the Amazon and in the Communication networks, this proportion is respectively 76.24 and 71,5 %. This result suggests that frequent conceptual links tend to be very homogeneous, since equivalent percentages of nodes belonging to a same community are found at the both sides of the pattern. In the EpiSims network, only 15.15 % of the values obtained are less than 0.5. In the Amazon and the Communication networks, this percentage is respectively 6,56 and 10,62 %.

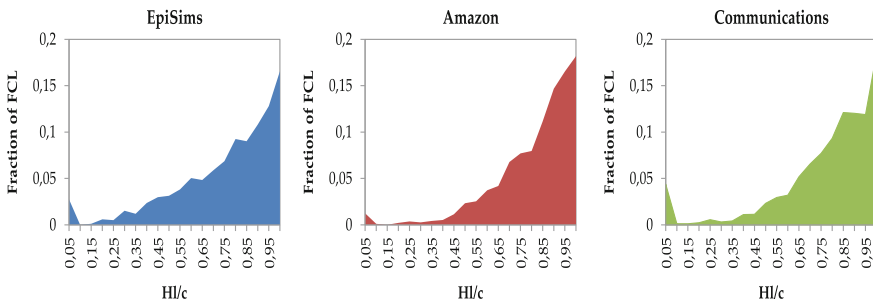


Fig. 10 Distribution of homogeneity rate $H_{l/c}$ of a community into a frequent conceptual link

Table 4 Examples of interesting patterns regarding $H_{l/c}$ (remind support of FCL: $\beta = 0.1$)

Network	Frequent conceptual link	Community	$H_{l/c}$
EpiSims	((*;1;*,*,*,*), (*;2;*,*,3;*))	13	1.0000
	((*;1;*,*,*,*), (*;2;1;*,*,*))	23	0.9995
	((*;2;2;*,*), (*;1;*,*,*,*))	7	0.9962
Amazon	((*,*,*,*,General;*,*), (Book;*,*,*,*,0;*))	22	0.9999
	((Book;*,*,*,*,*,*), (*;*,*,*,*,0;*))	27	0.9994
	((Music;*,*,*,*,*,*), (Book;*,*,*,*,0;*))	37	0.9989
Communication	((*,*,0;*,*,*,*), (*;*,1;*,*,0;*))	26	0.9998
	((Guadeloupe;*,0;*,1;*,*), (Guadeloupe;*,*,*,*,*))	25	0.9984
	((Guyane;*,*,*,*,*), (Guyane;*,1;*,*,*))	32	0.9958

conceptual links tend to connect nodes that belong to same communities. Moreover, this result demonstrates that a part of the intra-community links into a social network may be involved in a frequent conceptual link.

For example, the first line of the Table 4 provides relevant knowledge: First, ((*; 1; *, *, *, *), (*; 2; *, *, *; 3; *)) is a frequent conceptual link, i.e. at least 10 % of the links of the network connect men (*; 1; *, *, *, *) to women who have between 4 and 5 contacts (*; 2; *, *, *; 3; *) (we remind that the minimum link support threshold was set at 0.1). Second, in each group, the percentage of nodes that belong to community 13 is exactly the same.

6 Conclusion

In this paper, we have addressed the problem of clustering from social networks. Unlike traditional approaches that focus separately on the design of new patterns of knowledge suited to the context or the optimization of existing algorithms, we have adopted in this work another point of view by focusing on the fusion of patterns and the useful knowledge emerging from such fusions. For this purpose, we have focused on two network clustering approaches extracting two kinds of knowledge: (i) the clustering of nodes through the identification of communities and (ii) the clustering of links through the search for frequent conceptual links. The main contributions of the paper can be summarized as follows.

- We have formally described the concepts of communities and frequent conceptual links and discussed both the problematic and the useful knowledge resulting from their fusion.
- We have proposed a set of measures, based on the notion of homogeneity, that aim to evaluate the amount of shared information between these patterns when they are extracted from a same network.

- Finally, we have applied these measures to three datasets: a proximity-based network, a product co-purchasing network and a phone call network. The results obtained have demonstrated the interest of the approach proposed since very interesting merged knowledge have been identified on each network.

This work demonstrates the interest to consider simultaneously several sources of knowledge. In future works we plan to extend the approach to other network mining methods.

More generally, this work also raises a variety of questions in terms of visualization, extraction algorithms and resulting meaning. For instance in our future works, we plan to propose more complete representations of networks combining into single visualizations several kinds of knowledge. Another interesting track should be to propose optimized algorithms able to extract in one run several kinds knowledge.

References

- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487–499).
- Barrett, C. L., Bisset, K. R., Eubank, S. G., Feng, X., & Marathe, M. V. (2008). Episimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*.
- Blondel, V., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008.
- Croft, D. P., James, R., & Krause, J. (2008). *Exploring animals social networks*. Princeton: Princeton University Press.
- El Gamal, A. & Kim, Y.-H. (2011). *Network information theory*. Cambridge: Cambridge University Press.
- Fortunato, S. (2009). Community detection in graphs. *Physics Reports*, 486, 75–174.
- Ganter, B., Stumme, G., & Wille, R. (2005). Formal concept analysis, foundations and applications. *Lecture Notes in computer science* (Vol. 3626).
- Getoor, L., & Diehl, C. P. (2005). Link mining: A survey. *Physics Reports*, 7, 3–12.
- Kumpula, J. M., Saramäki, J., Kaski, K., & Kertész, J. (2007). Limited resolution in complex network community detection with potts model approach. *Physics Reports*, 56(1), 41–45.
- Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78, 046110.
- Lehmann, S., & Hansen, L. K. (2007). Deterministic modularity optimization. *Physical Review E*, 60(1), 83–88.
- Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web*, 1.
- Mangiameli, P., Chen, S. K., & West, D. (1996). A comparison of som neural network and hierarchical clustering methods. *European Journal of Operational Research*, 93(2), 402–417.
- Manyika, J., et al. (2011). Big data: The next frontier for innovation, competition, and productivity.
- Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577–8582.
- Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9), 2658–2663.

- Riadh, T., Le Grand, B., Aaufaure, M., & Soto, M. (2009). Conceptual and statistical footprints for social networks' characterization. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (p. 8). ACM.
- Scott, J. (2011). Social network analysis: Developments, advances, and prospects. *Proceedings of the National Academy of Sciences of the United States of America*, 1(1), 21–26.
- Stattner, E. (2014). Link formation in a telecommunication network. In *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1–9). IEEE.
- Stattner, E. & Collard, M. (2012a). Frequent links: An approach that combines attributes and structure for extracting frequent patterns in social networks. In *16th East-European Conference on Advances in Databases and Information Systems*.
- Stattner, E. and Collard, M. (2012b). Social-based conceptual links: Conceptual analysis applied to social networks. In *International Conference on Advances in Social Networks Analysis and Mining*.
- Stattner, E. and Collard, M. (2013). Towards a hybrid algorithm for extracting maximal frequent conceptual links in social networks. In *IEEE International Conference on Research Challenges in Information Science* (pp. 1–8).
- Steinhaeuser, K., & Chawla, N. V. (2010). Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31, 413–421.
- Yang, J., McAuley, J., & Leskovec, J. (2013). Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining (ICDM)* (pp. 1151–1156). IEEE.
- Yoon, S.-H., Song, S.-S., and Kim, S.-W. (2011). Efficient link-based clustering in a large scaled blog network. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, ICUIMC 2011* (pp. 71:1–71:5). New York: ACM.
- Zhou, Y., Cheng, H., & Yu, J. (2009). Graph clustering based on structural/attribute similarities. *Pattern Recognition Letters*, 2(1), 718–729.