Assessing group cohesion in homophily networks

Benjamin Renoust, Guy Melançon, Marie-Luce Viaud

Assessing group cohesion in homophily networks

Benjamin Renoust*†, Guy Melançon*, Marie-Luce Viaud†
*CNRS UMR 5800 LaBRI, INRIA Bordeaux Sud-Ouest
Campus Université Bordeaux I
Talence, France
{Benjamin.Renoust, Guy.Melancon}@labri.fr
†Institut National de l'Audiovisuel (INA)
Paris, France
mlviaud@ina.fr

***Abstract*—The analysis and exploration of a social network depends on the type of relations at play. Borgatti had proposed a type taxonomy organizing relations in four possible categories. Homophily (similarity) relationships form an important category where relations occur when entities of the network link whenever they exhibit similar behaviors. Examples are networks of co-author, where homophily between two persons follows from co-authorship; or network of actors having played under the supervision of the same movie director, for instance. Homophily is often embodied through a bipartite network where entities of a given type $\mathscr{A}$ (authors, movie directors) connect through entities of a different type $\mathscr{B}$ (papers, actors). A common strategy is then to project this bipartite graph onto a single-type network with entities of a same type $\mathscr{A}$, possibly weighting edges based on how the type $\mathscr{A}$ entities interact with the type $\mathscr{B}$ entities underlying the edge. The resulting single-type network can then be studied using standard techniques such as community detection using edge density, or the computation of various centrality indices. This paper revisits this type of approach and introduces a homogeneity measure inspired form past work by Burt. Two entities of type $\mathscr{B}$ interact when they both induce a same edge between two entities of type $\mathscr{A}$. The homogeneity of a subgroup thus depends on how intensely and how equally interactions occur between entities of type $\mathscr{B}$ giving rise to the subgroup. The measure thus differentiates between subgroups of type $\mathscr{A}$ exhibiting similar topologies depending on the interaction patterns of the underlying entities of type $\mathscr{B}$. The method is validated using two widely used datasets. A first example looks at authors of the IEEE InfoVis Conference (InfoVis 2007 Contest). A second example looks at homophily relations between movie actors that have played under the direction of a same director (IMDB).**

I. Introduction

The analysis and exploration of a social network depends on the type of relations at play. Borgatti [4] had proposed a type taxonomy organizing relations in four possible categories. The first type is *homophily* (also referred to as *similarity*) where connected members exhibit similar attributes such as membership in a club or interest group [15]. These types of ties do not represent actual social ties themselves, but might lead to a higher probability of a tie to develop between the members sharing similar attributes. Examples are networks of co-author, where homophily between two persons follows from co-authorship; or network of actors having played under the supervision of the same director, for instance.

The second type of ties are social relationships that can be affective relationships, friendship, ..., usually spanning over a long period of time. The third type captures joint interactions usually observed through discrete events such as calling each other or travelling together. The last type of ties describes flow (tangible or intangible) between entities (migrants moving between places, passengers between airports, etc.).

This paper focuses on networks induced from homophily relations,often embodied through a bipartite network where entities of a given type $\mathscr{A}$ (authors; movie actors) connect through entities of a different type $\mathscr{B}$ (papers; directors). Guillaume and Latapy [9] advocate bipartite graphs as being *universal* models for complex networks, hence offering additional motivations to use of these graphs to describe homophily relations.

When dealing with bipartite graphs, a common strategy is to project them onto a single-type network with entities of a same type $\mathscr{A}$, possibly weighting edges based on how the type $\mathscr{A}$ entities interact with the type $\mathscr{B}$ entities underlying the edge. The resulting single-type network often tends to have high edge density, with an increased propensity to contain cliques (depending on the affiliation data used to build the bipartite graph) [9]. It may nevertheless be studied using standard techniques such as community detection using edge density, or the computation of various centrality indices.

Referring to the work of Manski [14], we take the notion of a *group* as a central paradigm guiding the analysis of homophily networks. Numerous authors have indeed confronted homophily to many social behaviors or phenomenon (influence, contagion, information diffusion, e.g.) [1], [2], [19] questionning Manski's *group effect* as the driving force explaining the observed phenomenon.

This paper introduces a *homogeneity* measure inspired from past work by Burt [5] as a mean to inspect and assess group cohesion in a homophily network modeled as a bipartite graph. Two entities of type $\mathscr{B}$ interact when they both induce a same edge between two entities of type $\mathscr{A}$. When inspecting a subgroup of entities of type $\mathscr{A}$ (in the single-type network), we evaluate its homogeneity by measuring how intensely and how equally interactions occur between entities of type $\mathscr{B}$ giving rise to the subgroup. The measure differentiates subgroups of type $\mathscr{A}$ exhibiting similar topologies depending on the interaction patterns of the underlying entities of type $\mathscr{B}$.

Fig. 1 underlines the "nuance" we wish to bring into the analysis of homophily networks. Consider authors with attributes $A,\ldots,E$ in two different manners as shown in the Figure. Authors in the square node graphs (left) are linked by

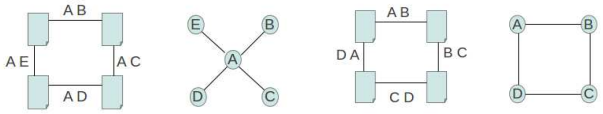


Fig. 1. An example underlining the "nuance" we emphasize by looking at how type $\mathscr{B}$ entities interact. In both figures, the square node graph (left) link type $\mathscr{A}$ entities (authors, movie directors, e.g.) whenever they are linked to a same entity of type $\mathscr{B}$ (keywords, movie actors, e.g.). Entities of type $\mathscr{B}$ appear as labels on induced links. The round node graph (right) describes how type $\mathscr{B}$ entities interact, that is when they co-occur as labels on an edge. The type $\mathscr{B}$ interaction network clearly distinguishes the two situations, whereas the projected single-type $\mathscr{A}$ networks show identical topologies.

an edge whenever they share an attribute. Observe that in both situations the pairwise "distance" between authors is the same, because any two authors share exactly two attributes, ending in identical topologies. As a consequence, based on pairwise distance, these two groups are somehow equivalent.

Now, consider the circle node graphs (right) describing how attributes interact *within the whole group* of actors. Clearly, *all* actors having attribute *A* gives this attribute a central position. The second situation is much more balanced (although attributes do not mix as intensely as they could). This small example points at situations where the analysis may be mislead when solely inspecting the single-type author network.

Groups can be computed using a variety of methods, from data clustering to community detection.Although advances have been made on that front in the past decades [11], [18] no algorithm or solution imposes itself as being superior in all situations. Understanding and validating the output of these algorithms is a challenging analysis task.

Our paper contributes an approach designed to help users evaluate the reliability of a proposed group structure. Because homophily (similarity) of actors is most often measured based on co-occcurences of attributes, we provide a mean to simultaneously work on the actor network *and* the attibute interaction network derived from the original homophily bipartite graph. The notion of a group here depends on the context: it may be a cluster computed from any algorithm, or a subset of authors selected by a user, for instance.

Our work exploits two statistics computed on any group of authors indicating the overall cohesion of the group measured through the intensity and homogeneity of interactions of their co-occurring attributes. Exploring the network, selecting a group or subset of co-occurring attributes and getting feedback on internal homogeneity, analysts can validate the model implicitly supported by the grouping procedure.

Our method has been validated based on two widely used datasets. A first example looks at homophily relations between movie directors that have directed the same movie actors (IMDB). A second example looks at co-authors of the IEEE InfoVis Conference where homophily is defined in terms of topics (keywords) associated with co-authored papers.

Related work. Bipartite graphs form an important modeling tool in social network analysis, supporting two-mode concepts [3]. They form an important analytical artifact to study homophily relations [8], and were even claimed as universal models for complex networks [9]. The literature covers a wide variety of approaches dealing with different properties of bipartie graphs and homophily networks. An optional but common strategy consists in projecting the graph inducing relationships between entities of a same type (see [10], [17], [20], for instance), with the obvious disadvantage of containing lots of cliques.

Because of their wide applicability and because they also offer a straightforward graphical representation of the data, bipartite graphs have been recently used in the design of a website traffic analysis system [6]. Finally, Kaski *et al.* [12] studied homophily in gene networks (similarity in gene expressions) in bio-informatics with emphasis on the trustworthiness of similarities, which places it close in spirit to our work.

II. Homophily, attributes and entanglement index

This section takes a closer look at homophily networks and describes the general framework we use.

As we shall see, group cohesion is easier to achieve with smaller groups. Inspecting a group, in an effort to understand why and how cohesion is embodied in the group certainly requires to be validated based on user knowledge. This only makes sense when conducted on small scale groups, gathering hundreds of nodes at most.

Simple questions come to mind when inspecting a group, such as "*How can we assess a group of actors really forms a cluster*?" "*How can we make sure all actors of a cluster really belong to it*?" "*Should we suspect the group to contain marginal (outlier) actors*?", etc.

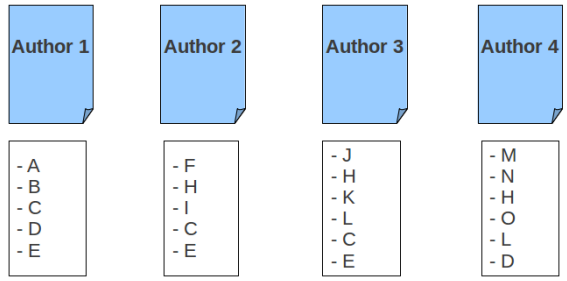
A central ingredient we used to answer these questions is a set of metrics that capture the intensity and homogeneity of interaction between attributes in a group of actors. These metrics can be viewed as an aid to assess of the internal homogeneity of a group.

A. Interaction network

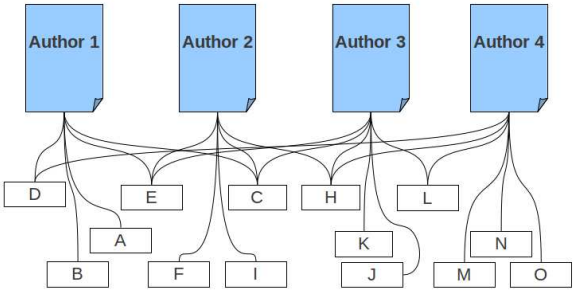
Our starting point is a set of actors $\mathscr{A}^*$ with associated attributes $A,B,C\ldots\in\mathscr{B}^*$, as shown in Fig. 2 (a). Most techniques use vector of attributes to compute distances between actors and infer semantically close groups of actors.

The inspection of a group of actors and associated attributes raises several questions. It might be important to know whether attributes equally map to *all* actors in the group, for instance. Conversely, a *misleading transitivity effect* may be suspected to take place. Indeed, we may have attributes t,t' co-occurring between actors a and a', and attributes t',t'' co-occurring between actors a' and a'', may lead one to believe that attributes t,t',t'' simultaneously co-occur between all three actors a,a',a''.

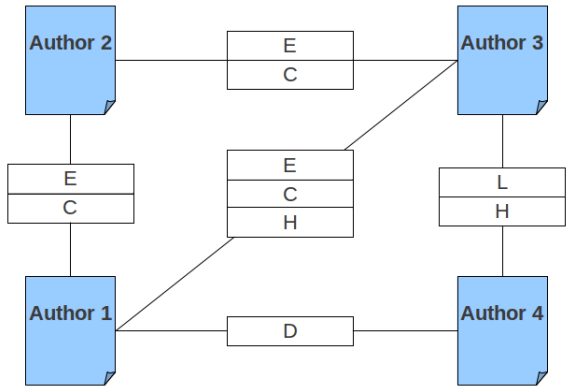
We addressed this issue by looking at how well attributes mix within a group. This is accomplished using the *entanglement index* computed for each attribute t, measuring how homogeneously an attribute co-occurs with all other attributes in a group of actors. Global homogeneity, at the group level, is then computed from the individual attribute entanglement indices. For instance, optimal homogeneity is reached whenever attributes have equal entanglement indices. This is the case



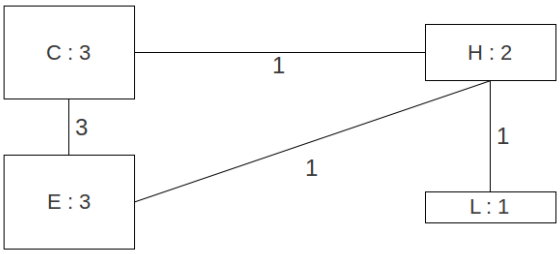
(a) Actors (authors) $\mathscr{A}^*$ writing papers with keywords $\mathscr{B}^*$



(b) Bipartite graph $G = (\mathscr{A}^* \cup \mathscr{B}^*, E)$



(c) Actor-actor graph $G_{\mathscr{A}} = (\mathscr{A}, E_{\mathscr{A}})$



(d) Attribute (keywords) interaction network $G_{\mathscr{B}} = (\mathscr{B}, E_{\mathscr{B}})$

Fig. 2. The initial data is formed of actors (authors) having associated attributes $A, B, C, \ldots$ (a) (keywords indexing papers). This situation is modeled by a bipartite graph linking actors to attributes (b) (authors having co-published papers on a given topic – keyword, see section III-B). We then consider the projected actor-actor network with attributes as edge labels (c) and derive the resulting *attribute interaction network* (d).

when all actors have the exact same associated attributes, and that all attributes equally co-occur within actors.

We now provide more details on these indices. The data can be usefully modeled as a bipartite actor-attribute network $G = (\mathscr{A}^* \cup \mathscr{B}^*, E)$ with edges $a - t$ whenever actor a has associated attribute t (see Fig. 2 (b)). Two other networks are derived from the actor-attribute network, namely an actor network $G_{\mathscr{A}}$ and an attribute interaction network $G_{\mathscr{B}}$. A actor network is usually built from the actor-attribute network by projecting paths $a - t - a'$ (linking actors $a, a' \in \mathscr{A}$ through attribute $t \in \mathscr{B}$) onto an edge $a - a'$ directly linking actors. We also need to store the attribute t as a label for the edge $a - a'$. Edges in $G_{\mathscr{A}}$ are thus labelled by subsets of attributes (all attributes $t, t', \ldots$ collected from triples $a - t - a'$, $a - t' - a'$, $\ldots$).

Because we are focusing on actor group cohesion and on attribute co-occurrence, we filtered out some of the edges. Loops were discarded, as well as edges $a - a'$ inferred from a single attribute t to obtain the actor network $G_{\mathscr{A}} = (\mathscr{A}, E_{\mathscr{A}})$. We moreover only kept edges $a - a'$ inferred from at least two different attributes t, t'. The resulting network is shown in Fig. 2 (c).

The attribute interaction network $G_{\mathscr{B}} = (\mathscr{B}, E_{\mathscr{B}})$ is built from attributes t that co-occur at least once with another attribute t' (through at least two actors). That is, there must exist at least two paths $a - t - a'$ and $a - t' - a'$ to infer the edge $t - t'$ in $E_{\mathscr{B}}$. Note that this network is *not* obtained by projecting paths $t - a - t'$ onto edges $t - t'$. For instance, $E_{\mathscr{B}}$ does not contain edges connecting descriptors that only concern a single actor. The attribute interaction network is a central artifact in computing group homogeneity.

Edges $t - t'$ moreover carry weights $n_{t,t'}$ indicating how often attributes co-occur between actors in the considered group. We also define $n_{t,t}$ to count the number of edges in $E_{\mathscr{A}}$ carrying the attribute t. The matrix $N_{\mathscr{B}}$ collecting all these $n_{t,t'}$ entries gives rise to another matrix $C_{\mathscr{B}}$ filled with ratios $c_{t,t'} = n_{t,t'}/n_{t',t'}$. The value $c_{t,t'}$ may be viewed as computing the (conditional) probability that an edge be of type t *given* it is of type t', while $c_{t,t} = n_{t,t}/N$ is the proprtion of edges carrying attribute t among all N edges in $G_{\mathscr{A}} = (\mathscr{A}, E_{\mathscr{A}})$.

Consider the example in Fig. 2. Starting from authors (actors) $a \in \mathscr{A}^*$ having published papers on topics $t \in \mathscr{B}^*$ (attributes), we build a bipartite graph where authors a, a' link through topic b whenever a and a' have co-authored a paper on topic b (Fig. 2 (b)). A single-type graph is obtained by inducing edges between authors labeled with topics (Fig. 2 (c)). The resulting attribute interaction network directly linking topics in shown in Fig. 2 (d). The matrices $N_{\mathscr{B}}$ and $C_{\mathscr{B}}$ (built over attributes C, E, H and L) then read:

$$N_{\mathscr{B}} = \begin{bmatrix} 3 & 3 & 2 & 0 \\ 3 & 3 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad C_{\mathscr{B}} = \begin{bmatrix} 0.3 & 1.0 & 1.0 & 0 \\ 1.0 & 0.3 & 0.5 & 0 \\ 0.3 & 0.33 & 0.2 & 1.0 \\ 0 & 0 & 0.5 & 0.1 \end{bmatrix}$$

B. Entanglement index

Inspired by Burt's work on relations ambiguity in multiple networks [5], we now wish to compute the *entanglement index* for each attribute, ,measuring how much an attribute t contributes to the overall cohesion of an actor group.

The entanglement value of an attribute t is reinforced through interactions with other highly entangled attributes. Denote by λ the maximum value among entanglement indices λ_t of attributes $t \in \mathscr{B}$. In other words, the entanglement index of attribute t is a fraction of λ, namely $\lambda_t = \gamma_t \cdot \lambda$. Having a probabilistic interpretation of the matrix entries $c_{t,t'}$ in mind, we can thus postulate the following equation which defines the values γ_t.

$$\gamma_{t'} \cdot \lambda = \sum_{t \in T} c_{t,t'} \gamma_t \tag{1}$$

The vector $\gamma = (\gamma_t)_{t \in T}$ collecting values for all attributes t, thus forms a *right* eigenvector of the transposed matrix C_I', as Eq. (1) gives rise to the matrix equation $\gamma \cdot \lambda = C_I' \cdot \gamma$. The maximum entanglement index thus equals the maximum eigenvalue λ of matrix C_I'.

The actual entanglement index values λ_t are of lesser interest; we are actually interested in the relative γ_t values. Furthermore, we shall see how the entanglement vector γ and eigenvalue λ can be translated into a network measures to help understand entanglement in a group of actors.

C. Homogeneity and intensity

Let us see how our indices can help measure global *entanglement intensity and homogeneity*. The topology of the attribute interaction network $G_{\mathscr{B}} = (\mathscr{B}, E_{\mathscr{B}})$ provides useful information about how attributes contribute to the overall cohesion among actors of a group. The focus here is on interactions among attributes and aims to reveal how cohesive the group of actors is considering this set of attributes.

The archetype of an *optimally cohesive actor group* is when all actors have the exact same associated attributes. In that case, the graph $G_{\mathscr{B}} = (\mathscr{B}, E_{\mathscr{B}})$ then corresponds to a *clique*. As a consequence, all matrix entries $n_{t,t'}$ coincide, so all entries in matrix C_I equal 1. The maximum eigenvalue of C_I' then equals $\lambda = |\mathscr{B}|$, and all γ_t coincide. That is, all attributes indeed contribute, and they all contribute equally to the overall actor group cohesion. The Perron-Frobenius theory of nonnegative matrices [7, Chap. 2] further shows that $\lambda = |\mathscr{B}|$ is the maximum possible value for an eigenvalue of a non-negative matrix with entries in $[0,1]$.

The Perron-Frobenius holds for irreducible matrices, that is when the graph $G_{\mathscr{B}}$ is connected. Hence, the connected components in $G_{\mathscr{B}} = (\mathscr{B}, E_{\mathscr{B}})$ must be inspected independently. When When the matrix $C_{\mathscr{B}}$ is irreducible, the theory of non-negative matrices tells us that it has a maximal real positive eigenvalue $\lambda \in \mathbb{R}$, and that the corresponding eigenvector γ has non-negative real entries [7, Theorem 2.6]. We hereafter assume that C_I is irreducible.

Inspired from the clique archetype of an optimally cohesive actor group, we wish to compute a cohesion index at the actor group level. We already know that the eigenvalue is bounded above by $|\mathscr{B}|$, so the ratio $\frac{\lambda}{|\mathscr{B}|} \in [0,1]$ measures how intensely interactions take place within the actor group. This ratio thus provides a measure for cohesion *intensity* among all actors *with respect to attributes in* $\mathscr{B}$.

We also know that the clique situation with equal $c_{t,t'}$ matrix entries leads to an eigenvector γ with identical entries. This eigenvector thus spans the diagonal space generated by the diagonal vector $1_{\mathscr{B}} = (1,1,\ldots,1)$. This motivates the definition of a second measure providing information about how homogeneously cohesion distributes among attributes. We may indeed compute the cosine similarity $\frac{\langle 1_{\mathscr{B}}, \gamma \rangle}{||1_{\mathscr{B}}|| \cdot ||\gamma||} \in [0,1]$ to get an idea of how close the actor group is to being optimally cohesive. We will refer to this value as cohesion *homogeneity*.

A thorough study of the entanglement indices, and the homogeneity and intensity network indices is out of the scope of this paper (see [16]).

D. Analysis of homophily networks

Attributes entanglement indices, and the homogeneity and intensity measures may thus be used to inspect homophily networks and assess cohesion in subgroups of actors. These computational devices suggest to use a synchronized dual view of a homophily network splitted into two distinct but complementary networks: the networks of actors $G_{\mathscr{A}} = (\mathscr{A}, E_{\mathscr{A}})$ and the interaction network of attributes $G_{\mathscr{B}} = (\mathscr{B}, E_{\mathscr{B}})$.

Typically, when using a node-link view of these networks, the selection of a set of actors should automatically trigger the selection of the relevant attributes and compute the corresponding entanglement, homogeneity and intensity values. This is illustrated in Fig. 4), where a set of movie directors has been selected (top panel). Movie actors that played under their direction, here seen as attributes of movie directors, are highlighted (bottom panel). The corresponding homogeneity index, restricted to these four selected directors, are displayed as a background of the selection lasso, while the actual values are reported in a side panel. The size of movie actors nodes corresponds to their entanglement index: a larger node indicates a movie actor weighs more in bringing these movie directors together as a group.

III. Case studies

The use cases we describe in this section aim at showing how the entanglement indices, and the homogeneity and intensity indices of networks help users explore social networks and reason about the homophily content. The examples are designed to highlight different aspects of the exploration, each time underlining how the indices contribute to better understand the group structure of the homophily network.

Roughly speaking, the knowledge users gain after applying a grouping procedure (clustering, community detection) is that "*a group of actors*" share "*a list of attributes*". This is where the entanglement index enters the scene. What does "*a list of attributes*" really mean? Do all actors share all attributes? Do actors more or less split between attributes? What particular attribute(s) make(s) the split explicit? In other words, users must be able to elucidate to what extent, and possibly how/why, the group of actors form a more or less homogeneous unit.

A. IMDB

This first use case is built from the Internet Movie DataBase, a largely used dataset (*www.imdb.com*).Starting from the main actors in a chosen subset, we have additionally extracted the corresponding movie directors to form a bipartite network where directors connect to movie actors they have directed. Applying our methodology we compute (i) a movie

director network, where two directors connect when the set of actors they have directed share at least two actors, together with (ii) the corresponding actor interaction network. The data may thus be used to find homogenous subgroups of movie directors, those whose artistic signature rely on similar movie casts.

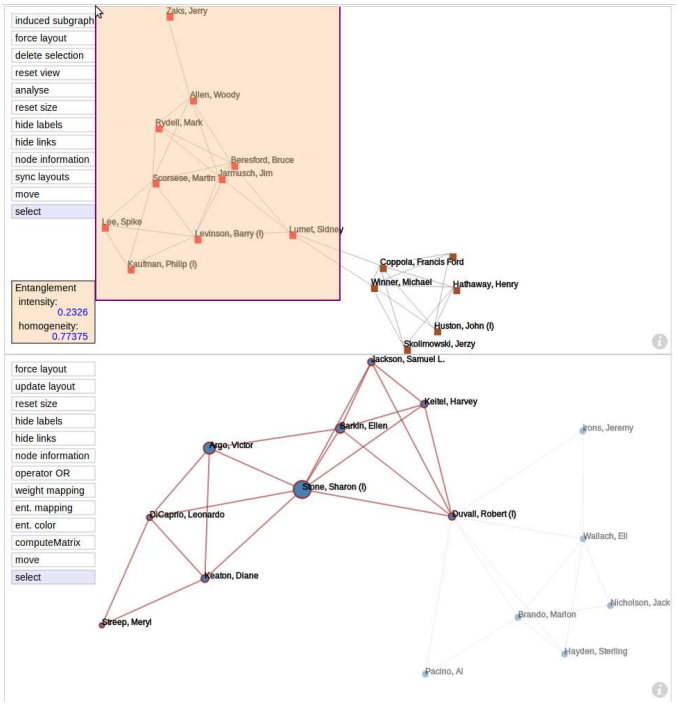


Fig. 3. IMDB - *directors* appear on top; *actors* on bottom. Selecting a group of directors highlights the corresponding actors, with node size mapped to their entanglement index. This group of directors shows low homogeneity and intensity. We can clearly see that the distribution of actors is unbalanced, partly because Sharon Stone plays by far a central role in the interactions between directors – the directors all have, at some point, directed her.

This first example gathers 15 actors and 16 directors (see Fig. 3). A low intensity and medium homogeneity, together with a loosely connected actor interaction network topology suggest that actors and directors roughly split into two communities. The director network has medium homogeneity that corresponds to a quite balanced distribution of actors among them. Homogeneity is not optimal: the directors did not individually direct *each* of these actors although, as a group, they did direct *all* of these actors. This indicates the need to dig further and try to "nuance" the homogenity of this group. Roughly speaking, low intensity is explained since most directors have directed only a small number of actors relatively to the whole set. As can be seen from Fig. 3 (bottom), the two communities of actors are connected through R. Duvall, and the two communities of directors are connected through S. Lurnet.

The "M. Brando" sub-community (bottom right) shows higher intensity (with homogeneity similar to the overall network). These actors appear as attributes for a subgroup of directors centred around F.F. Coppola, J. Huston, and others. Higher intensity means they played with many other actors under the direction of these movie directors.

The community of actors located in the top left part of the

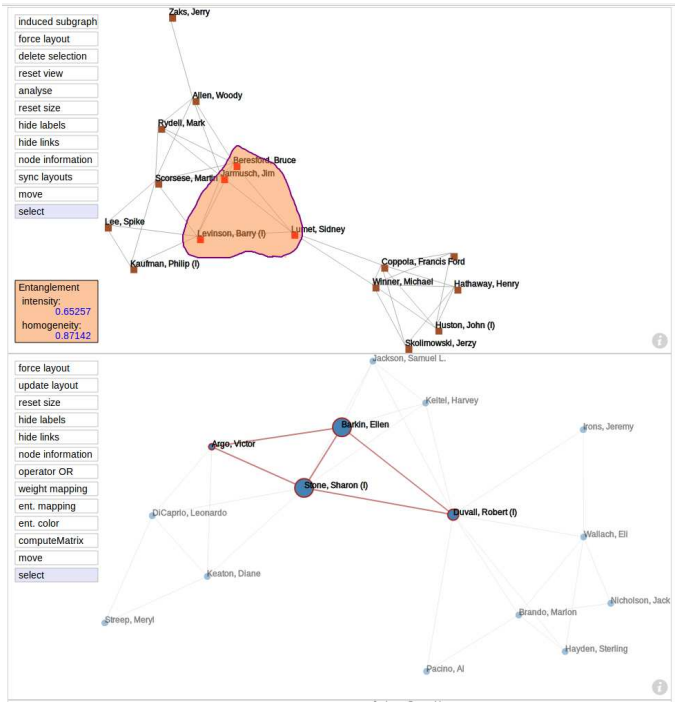


Fig. 4. A group of directors (top) and the corresponding actors they co-directed (bottom, highlighted) with node size mapped to their entanglement index. This clique of 4 directors shows higher homogeneity and intensity than the selected group on Fig. 3.

panel correspond to a different group of directors (connecting to the previous group through S. Lurnet). It gathers S. Lee, J. Jarmusch, M.Scorsese, W. Allen and others. This community has similar intensity but higher homogeneity when compared to the overall network. This means these actors have equal influence within this group and better capture altogether the artistic signature of these directors as a group.

The upper left subgroup in the director network (see Fig. 4) actually divides into three overlapping cliques. Two cliques reach maximal homogeneity and intensity (the exact same actors have all played under their direction). The third clique (B. Beresford, J. Jarmusch, B. Levinson, and S. Lurnet) – selected in the top panel of Fig. 4 – focuses on E. Barklin and S. Stone. It has lower homogeneity and intensity indices: they don't mix that well with the other actors.

This use case thus underlines the fact that although a group involve a well identified and distinct set of attributes (movie actors), the homogeneity of the considered group may rely only on a subset of these attributes. Clearly, group cohesion must not solely rely on the topology of the projected single-type network obtained from the original bipartite network.

B. InfoVis 2004 contest

Our second example brings data of a different nature, where topics (keywords) link to authors, showing that the notion of homogeneity can actually apply to a wide variety of entities.

We selected a subset of the InfoVis 2004 Contest dataset giving papers published in InfoVis between 1994 and 2004 [13]. The data we consider are authors indexed by keywords gathered from papers they published. We thus compute a

bipartite graph where authors link to keywords. To some extent, with respect to Borgatti's taxonomy of relations [4], this network could be considered as an *interaction* network since co-authorship indeed involves direct contact with collaborators.

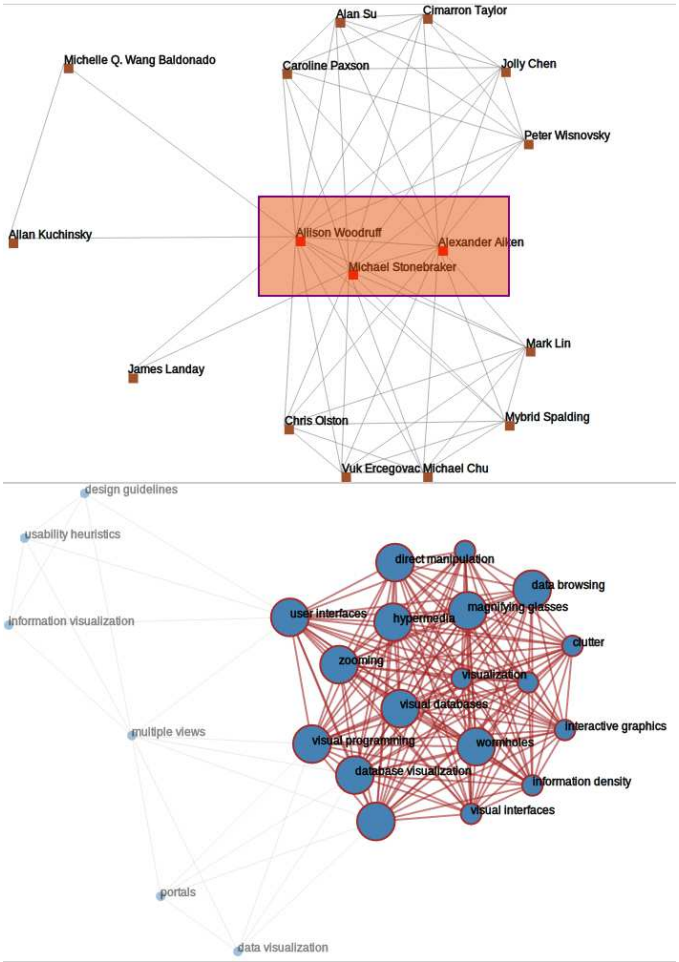


Fig. 5. The InfoVis 2004 Contest data gives rise to a topic (keywords) interaction network paired with an author social network. The three selected authors (top panel) hold a central position in the social network (top). Their co-publications cover a wide spectrum of topics as shows the clique of terms in the bottom image. Homogeneity, although good, is however not optimal: they did not pairwise co-published on all these topics. We may indeed suspect each of them to have different co-authors in the network.

When we consider authors and keywords, groups may form because authors are socially very close – working in same laboratories, graduated same institutions – or just formed an opportunistic association around trendy topics. We took this aspect in consideration by making sure that authors were connected through a keyword only when they indeed had co-published a paper on that topic – not just because they both had published a paper on that topic.

We show how our approach helps to solve two tasks of the InfoVis 2004 Contest: *Where does a particular author/researcher fit within the research areas? What, if any, are the relationships between two or more or all researchers?*

The author-keyword bipartite graph gives rise to a keyword interaction network $G_{\mathscr{B}}$ and an author social network $G_{\mathscr{A}}$. The full social network contains about 1000 authors and breaks into several connected components. We will focus on the component lead by Woodruff, Olston and Stonebraker (see [13, leftmost part of Fig. 4]) gathering 16 authors (see Fig. 5 – top).

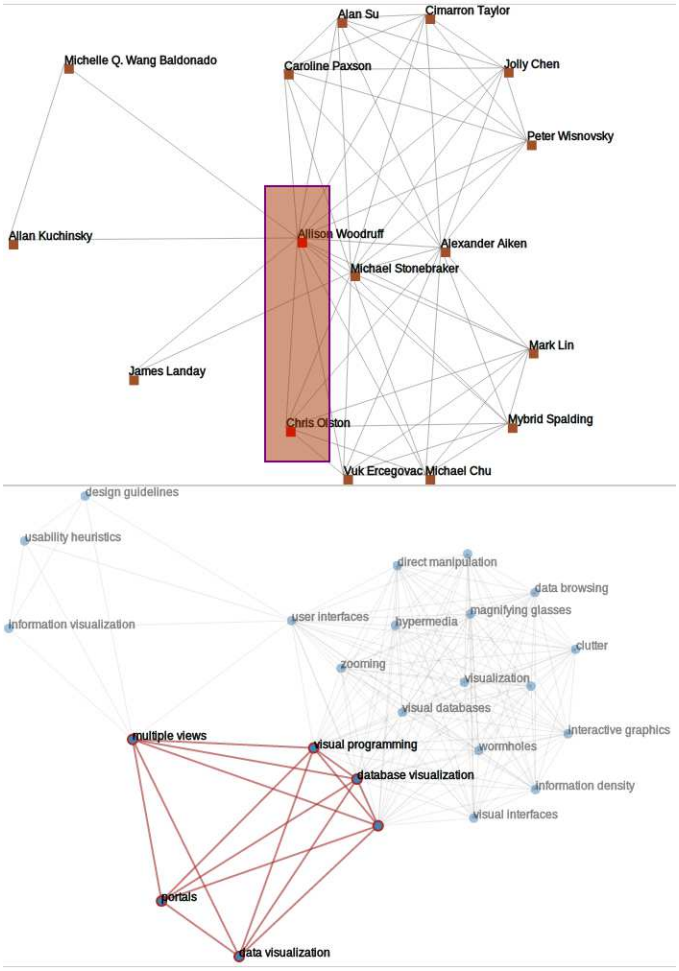


Fig. 6. Browsing around "obvious" sub-communities of authors, the keywords *portals* and *data visualization* never pop up. Directly selecting them in the keyword network brings two co-authors up front: Woodruff and Olston (top). Selecting these authors shows their common topics of interest to be marginally positioned with respect to the main clique (bottom).

The answer to the first question is straightforward. Selecting a single author, its associated keywords are highlighted while positioned in the context of neighbor topics. The social network displays the co-authors of any selected author.

Inspecting the whole network author by author is lengthy and tiresome and cannot reasonably be performed on larger networks. This brings us to the second task requiring a more elaborated exploration strategy. In our case, we may take benefit of the apparent community structure of the social network. Conversely, we may select a subset of keywords and look at authors who have published on these topics to see how homogeneous a community they form, for instance.

The topology of the author network (Fig. 5 – top) clearly shows three authors as central actors (A. Woodruff, M. Stonebraker and A. Aiken) at the intersection of two different cliques. Their associated keywords form a large clique covering a large part of the keyword network (Fig. 5 – bottom). The entanglement values (node sizes) widely vary among keywords

explaining why homogeneity is low, moreover suggesting that each of these three authors have her/his own set of topics.

Selecting the authors that are part of the top clique in the social network (*Paxson*, *Wisnovsky*, ...), *except* those central actors leaves us with a subset of authors with optimal homogeneity: they all co-published on the exact same topics. The same is true with the authors of the bottom clique (*except* the central authors – *Olston*, *Spalding*, ...).

We may also select two marginal authors sitting on the left side (Baldonaldo & Kuchinsky) and observe that they link to keywords located out of the "Woodruff clique" keyword subsets. Strikingly,none of these sub-communities seem to address the topics *portals* and *data visualization* located at the bottom left of the keyword network. Grasping these two keywords, we find that they solely concern Woodruff and Olston. Going back to the author network, selecting Woodruff and Olston we then see the additional topics these two authors have in common. Observe how these topics are marginally positioned with respect to the main clique (Fig. 6 – top).

This second use case pointed at fully homogeneous subgroups where authors have co-published papers on the exact same topics. This also suggest that the analysis may be conducted either from the actor (author) network *or* the attribute (keywords) network. Going back and forth between these two perspectives seems a fruitful strategy to get the most out of the entanglement index and the dual $G_{\mathscr{A}} - G_{\mathscr{B}}$ representation.

IV. Conclusion and future work

This paper addressed the issue of assessing cohesion in groups from homophily networks mixing actors and attributes into a bipartite graph. Our approach considers splitting the bipartite into two single-type networks used in conjunction when analyzing the homophile relations between actors. To answer this question, we have defined the entanglement index on attributes, together with the homogeneity and intensity indices computed on any subset of attributes.

These attributes can be used to question the homogeneity of a group, where optimal homogeneity requires that actors simultaneously involve the *exact same* attributes, and maximum intensity occurs when actors cover *all* available attributes. A group of lower or unbalanced homogeneity indeed requires more careful analysis, and typically leads to the discovery of subgroups or regions locally showing higher homogeneity.

The case studies clearly show the relevance of questioning the attribute homogeneity of actors to potentially confirm the community structure derived from edge density, for instance. They focused on small size examples for sake of readability. This limitation is but apparent, as using the interaction network occurs after actors have been indexed and grouped. Although a query might return hundreds (or thousands) of actors, we may expect the grouping procedure to form much smaller groups before closer examination occurs. We also suspect that larger samples gather larger attribute sets, typically leading to less tangled attribute interactions and less homogeneous groups.

Our second case study suggests our approach applies to other types of networks modeled using a bipartite graph, namely interaction relations. There even is a potential to extend our approach to the study of multivariate networks. Extending the methodology to weightened relationships is also an avenue we plan to explore. These are design choices we suspect may depend on the nature and/or on the size of the dataset.

We also plan to examine strategies to automatically identify attribute and actor subsets with optimal (or maximum) homogeneity and/or intensity, suggesting potential areas of interest in the network under study. These problems, however, will inevitably bring us to combinatorial optimization problems, and we may expect to have no choice but to rely on heuristics to avoid typical algorithmic complexity issues.

References

[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[3] S. P. Borgatti. Two-mode concepts in social network analysis. In R. A. Meyers, editor, *Computational Complexity - Theory, Techniques, and Applications*, pages 2912–2924. Springer, 2012.

[4] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

[5] R. Burt and T. Scott. Relation content in multiple networks. *Social Science Research*, 14:287–308, 1985.

[6] W. Didimo, G. Liotta, and S. A. Romeo. A graph drawing application to web site traffic analysis. *Journal of Graph Algorithms and Applications*, 15(2):229–251, 2011.

[7] J. Ding and A. Zhou. *Nonnegative Matrices, Positive Operators and Applications*. World Scientific, Singapore, 2009.

[8] D. Easley and J. Kleinberg. Networks in their surrounding contexts. In *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*, pages 77–106. Cambridge University Press, 2010.

[9] J.-L. Guillaume and M. Latapy. *Bipartite Graphs as Models of Complex Networks*, volume 3405 of *Lecture Notes in Computer Science*, pages 127–139. Springer, 2005.

[10] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2010.

[11] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[12] S. Kaski, J. Nikkila, M. Oja, J. Venna, P. Toronen, and E. Castren. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4(1):48, 2003.

[13] W. Ke, K. Borner, and L. Viswanath. Major information visualization authors, papers and topics in the ACM library. In *IEEE Symposium on Information Visualization 2004*. IEEE, 2004.

[14] C. F. Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993.

[15] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[16] G. Melançon, B. Renoust, and M.-L. Viaud. Mesurer la cohésion sémantique dans les corpus de documents. Technical Report RR-8075, INRIA Bordeaux Sud-Ouest, 2012.

[17] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[18] M. Plantié and M. Crampes. Survey on social community detection. In N. Ramzan, R. Zwol, J.-S. Lee, K. Clüver, and X.-S. Hua, editors, *Social Media Retrieval*, Computer Communications and Networks, pages 65–85. Springer, 2013.

[19] C. R. Shalizi and A. C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

[20] T. Zhou, J. Ren, M. Medo, and Y. Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):046115, 2007.