https://doi.org/10.2352/ISSN.2470-1173.2019.1.VDA-686

M-QuBE^3 : Querying Big Multilayer Graph by Evolutive Extraction and Exploration

Antoine Laumond, Guy Melançon, and Bruno Pinaud; Univ. Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France
Mohammad Ghoniem; Luxembourg Institute of Science and Technology (LIST)

Abstract

Although node-link representations of graphs are widespread and even sometimes preferred to other approaches, they suffer from obvious limitations when graphs become large or dense, inducing visual cluttering and impeding the traditional visual information seeking process. This article presents a new strategy of exploration particularly suitable when graphs are large and dense. Users iteratively drive the exploration through the visualization of small sub-networks of interest. Our technique is particularly useful with multilayer networks, where layers typically combine into a large and dense network. Our iterative exploration process called M-QuBE^3 *computes a score for each node of a graph based on structural and semantic information where more interesting nodes from a user point of view have higher scores. This in turn translates into a procedure to select sub-networks of interest. Within each sub-network, the user can select nodes to enhance the semantic context (and thus impact their interest score) and iteratively refine the exploration towards more relevant sub-networks. The* M-QuBE^3 *process natively handles multilayer network and allows the use of layers as a semantic apparatus when driving the navigation.*

Introduction

Let Estelle be a historian active in European Integration studies. As part of her research she often investigates the role of a public figure or organization in the European Integration process. Among other options, she has access to a large knowledge base of papers produced in-house by fellow researchers, citing source documents related to key persons, organizations, locations, etc. The source documents may be of various types including diplomatic correspondence, minutes of meetings, newspaper articles, cartoons, audio and video footage, etc. Since this knowledge base is a large collective creation, accumulated over many years, individual subject matter experts like Estelle can only have a partial knowledge of its contents. In addition, when Estelle conducts a new study, only a fragment of the entire knowledge base may prove relevant. She would also like to set some constraints such as striking a balance between the types of source documents used as references e.g. more diplomatic letters than newspaper articles. In the course of her investigation, Estelle will retain or discard certain documents or entities based on her expertise.

In brief, Estelle needs a way to discover and refine progressively the constituents of the story she will eventually tell, starting from a very partial list of relevant entities, with some constraints to meet regarding the inclusion and distribution of some entity types. For instance, in order to develop the impact of the Cold War on Europe, she starts from an element representative of this period (e.g. Glasnost, Perestroika, Cuba, Berlin Wall), then extends and expands the search from the elements she found.

In this simple description, the knowledge base comprises a very large number of entities and documents with numerous direct document-entity relationships, and much more derived entity-entity co-occurrence relationships e.g. person-person or person-location relationships. In Information Visualization terms, this is a case of large and dense graph with heterogeneous nodes and various types of links, as well as attributes associated with the nodes, e.g. a person's dates of birth and death, which may be described as a multilayer multivariate time-dependent graph often referred to as multilayer graphs [11]. While exploring and navigating through large information spaces is the main raison d'être of information visualization, large and dense network datasets are known to defeat well-established approaches such as the visual information seeking mantra [13] ("overview first, zoom and filter, then details on demand"). This is partly due to the fact that popular graph layouts (e.g. node-link) result in unwieldy overviews due to edge clutter, which precludes any further exploration [6, 7]. More importantly, the top-down approach assumed by this mantra contradicts the analytical methodology adopted in some fields (e.g. by historians), where the scope of investigation starts from a detail and expands gradually to include more and more elements. The Visual Analytics mantra [10] ("Analyze first, show the important, zoom, filter and analyze further, details on demand") introduces an automated analysis step at the onset of the analytical pipeline in order to qualify what is important to the user. This is useful when providing an exhaustive overview becomes infeasible due to the sheer size of the dataset (computing power) or to cognitive overload. In our practical case, this approach is limited because what is important for the experts is not initially fully specified, as in the previous example about the Cold War.

To take into account both the complexity of the data structure and the specific analytic workflow (expand from a detail) we propose in this paper a new technique called M-QuBE^3. It is an incremental and interactive method to visually mine digital cultural heritage data (structured in a multilayer network) based on node interest calculations to extract series of sub-networks of manageable size and increasing quality capturing the evolving criteria of the domain expert. Overall, the contributions of this paper are:

1. An incremental exploration mechanism which gradually increases the relevance of the visualizations based on node interest computation;
2. eScore: an iterative algorithm computing node interest based on an incremental node selection designed for multilayer graph;
3. M-QuBE^3, a combination of the two previous items, which was implemented and validated using expert feedback.

The rest of this paper is structured as follows. We first present related work, then we detail the M-QuBE^3 process by explaining its iterative operations and its interest-based score calculation. Next, two case studies are presented on historian data. Finally, we conclude after a discussion part.

Related Work

The challenge we take up in this work is to determine the most meaningful elements according to a personal query, which is what recommender systems do. Popular approaches are based on a ranking of the searched elements [1], using additional information from meta-data (themes, categories, etc.) [3]. In any case, user action is limited by the initial query. Contrary to recommender systems, one design rationale of M-QuBE^3 is constant interactions with users to let them refine and enhance the obtained recommendations. M-QuBE^3 moreover takes advantage of the network topology features, and multilayer structure of the data.

Quantifying user interest is the key to create an efficient way to rank network elements. The starting point is the founding work of Furnas [5]. Its most noteworthy example is a code editor organized in a tree, where the edited code portion is fully accessible, and placed in the context of the whole module to which it belongs. Code blocks farther away are then summarized as a single line or function header. Furnas' work consists in specifying a good notion of "distance" to decide what to display, in detail or in summary form. This notion of distance can be expressed as a score for estimating node interest in different domains such as ontologies [9] or trees [2]. Van Ham and van Wijk used this approach for navigation purposes in a tree describing a hierarchical ascending partitioning of the nodes of a graph (clustering) [8]. Van Ham and Perer proposed a generalization of the score to any graph, adding a semantic dimension [14]. Interest is determined from structural information of the graph (degree, centrality, etc.) and information relative to node attributes (keywords, tags, etc.). M-QuBE^3 generalizes and aggregates previous work by leveraging multilayer graphs, more faithfully capturing the complexity of the data than traditional single-layer graphs.

From a visualization point of view, existing multilayer graph visualizations, e.g. MuxViz [4], can benefit from the M-QuBE^3 process, as it proposes both an application adapted to these visualizations as well as a procedure of exploration in order to put the user in control of his navigation, a need previously identified by McGee et al. [12].

Evolutive Exploration Through Partial Views

As stated in the introduction, numerous node-link graph navigation methods do not allow efficient navigation and visualization for multilayer networks. Moreover, navigation can only be done iteratively to capture the workflow of domain experts. At any time in the visualization process, the analyst may keep refining the path she has followed so far, or decide to question it and start a new search path from scratch.

We introduce below the M-QuBE^3 process (for **M**ultilayer network: **Qu**erying **B**ig networks by **E**volutive **E**xtraction and **E**xploration) consisting in building a succession of sub-networks to mimic the workflow of the experts.

The general idea, as depicted in Fig. 1, consists at each iteration in using user inputs (keyword search or node selection) to compute a transient sub-network directly from the initial network such that it becomes increasingly pertinent to the user. Therefore, experts are able to explore and guide their exploration through a large network by simply analyzing a series of reduced sub-networks instead of the entire network.

The M-QuBE^3 process is split into three main phases (Fig. 2). The process starts with a keyword search (Panel A). Its results, i.e. the selected nodes, form the initial *focus set* (**A3**): a list of reference nodes that allow to define candidate nodes that may potentially be displayed in the extracted sub-network. A score is calculated for these candidate nodes by considering the semantic information of the dataset and the network structure in order to estimate their interest for the user (Panel B). From these scores, a ranking is made to determine a list of the most interesting nodes (chosen list) which are then used to extract the sub-network which is shown to the user (Panel C).

Because experts start with a detail and progress step by step, the M-QuBE^3 process has a similar workflow. The M-QuBE^3 process may be repeated at will in order to explore the data more and more deeply and with greater precision. The user interacts with the resulting sub-network at each iteration by selecting new nodes which they deem relevant. The nodes enrich the focus set and thus improve the next sub-networks. For this purpose, the focus set is kept throughout the series of sub-network extractions and used to compute the next sub-network. We detail below the different steps that make one iteration of the process.

Focus Selection (Fig. 2, Panel A)

First, the focus set needs to be initialized or updated. At the first iteration, the focus set (A3) is initialized by a keyword search (A1). Then, in each new iteration, the focus set is modified by adding or removing nodes based on user selection (A2). Neighbors of the focus set (A4) then compose the candidate list (B1) which contain the nodes considered for inclusion in the next extracted sub-network (C6) depending on the evaluation described in Panel B. This list of candidate nodes evolves throughout the process. The next phase is to estimate the user interest of this candidate list.

Interest Computation (Fig. 2, Panel B)

The interest phase computes a score (B7) which represents user interest for a given node x. The higher the score, the more likely it can be selected and shown to the user in the next sub-network (C6). This phase first consists in computing the interest score (eScore, B3) along with a position score (pScore, B4), which are combined to give a weighted score (B5). The final score is obtained by also taking into account the score of the neighbours of x (same process, B3' to B5') using a diffusion calculation (B6).

eScore computation (B3). The eScore is computed for all nodes in the candidate list (B1). Details of how it works are explained in the next section (eScore: interest estimation).

pScore computation (B4). Users interact with the process by selecting nodes at each iteration (Panel A). We assume that nodes close to a selected node, in the sense of geodesic distance, have more chances to be considered interesting by the user. User selection thus constitute what we call a focal zone and a node included or close to this zone is weighted positively (see weighted score

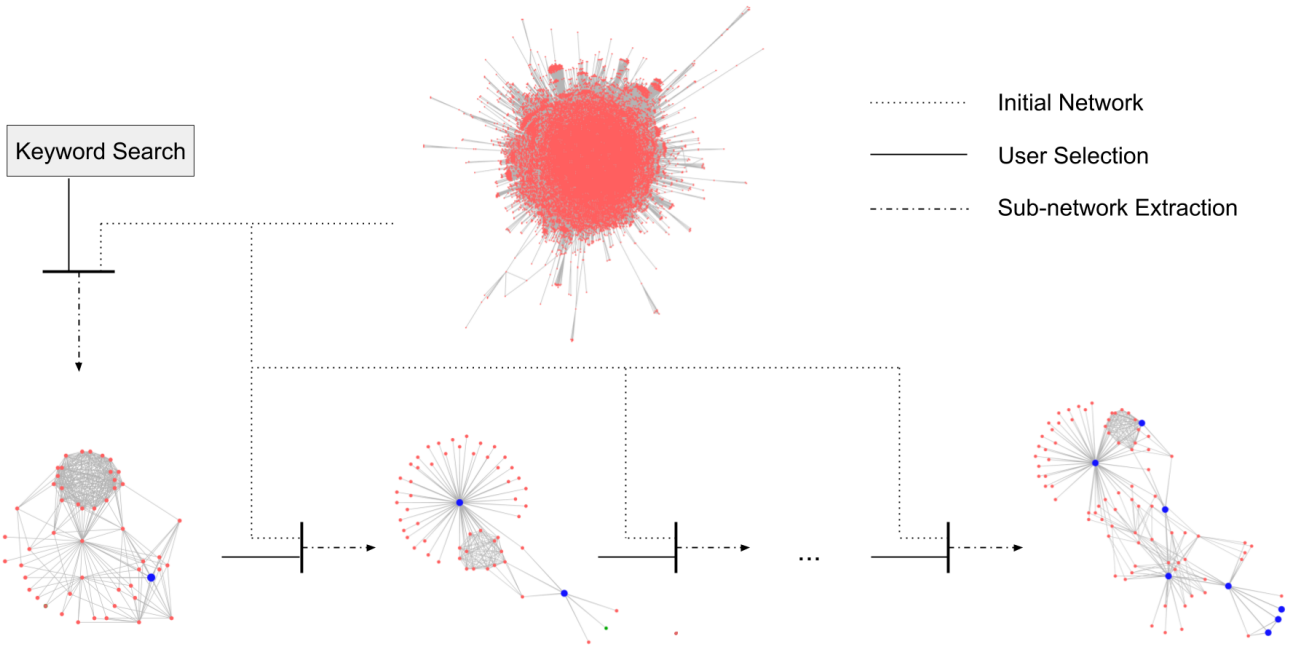


Figure 1: M-QuBE[3] from an expert point of view: Creation of a series of sub-networks based on a measure of interest. The "hairball" is the initial network and is not (intended to be) displayed to the user. However, the user selects one or more nodes in it using keyword search among node names and attributes. The first sub-network is extracted based on the nodes selected by the keyword search. Then, the user selects interesting nodes (blue nodes) in the sub-network to create another one according to the new selection. The process continues until the user is satisfied with the sub-network at hand.

(B6) computation below). To this end, a centroid-based function is used to compute the average distance between the evaluated node x and the nodes of the focus set, thus determining its position in the focal zone. This function is defined as:

$$C(x,Y) = \frac{\sum_{y \in Y} d(x,y)}{|Y|}$$

with Y the focus set and d a distance function. The most relevant function for d regardless of the context is often the shortest path between two nodes. Euclidean distance may also be used but the coordinates of the nodes given by a layout algorithm have to make sense, which requires some work on the network layout in the first place.

$pScore(x,Y)$ is the normalized distance between x and Y:

$$pScore(x,Y) = 1 - \frac{C(x,Y) - c_{min}}{c_{max} - c_{min}}$$

with c_{min} and c_{max} respectively the maximum and minimum value of C in the network. The normalization is necessary to rank nodes afterwards (C1) because nodes in the candidate list have obviously different neighborhoods.

weighted score (B5). Both *pScore* and *eScore* are combined into the weighted score (*wScore*) by the following function:

$$wScore(x,Y) = (1-w) \times eScore(x|Y) + w \times pScore(x,Y)$$

with w a constant on $[0;1]$ set by the user to give more or less importance to the focal zone. wScore represents the estimated interest of a node by taking into account both the semantics (eScore) and the structural information of the network (pScore).

Diffusion (B6). The computation of node interest ends with the diffusion phase. A possible problem of this process is the same as that encountered by van Ham and Perer [14]. The list of candidate nodes extends iteratively like a greedy algorithm: when a node is selected and enters the chosen list, its neighboring nodes are added to the candidate list. However, if a very interesting node (a node with a high score) is surrounded by nodes with a low score, the iterative algorithm may never select it (since its neighbors may never be selected). If we want to avoid these isolated local extrema, we proceed to a diffusion of the score of interest.

The solution consists for each interesting node to diffuse a part of its interest to its neighborhood. To do this, the users select a degree dif of diffusion. The higher dif is, the wider the diameter of the extracted network may be (If dif is zero, then the diffusion mechanism is not used).

To make a diffusion of degree dif of a node, it is then necessary to calculate the score of non-candidate nodes at distance dif from it (B2). Then, once the score is calculated for all the required nodes (B3',B4',B5'), each node gains a percentage of the weighted score of the most interesting node (the node the maximum score) at distance dif or less (B6). This percentage is also set by the user. The higher it is, the more similar the node scores become. This optional mechanism can potentially improve the relevance of the sub-networks of interest we get. However, a high

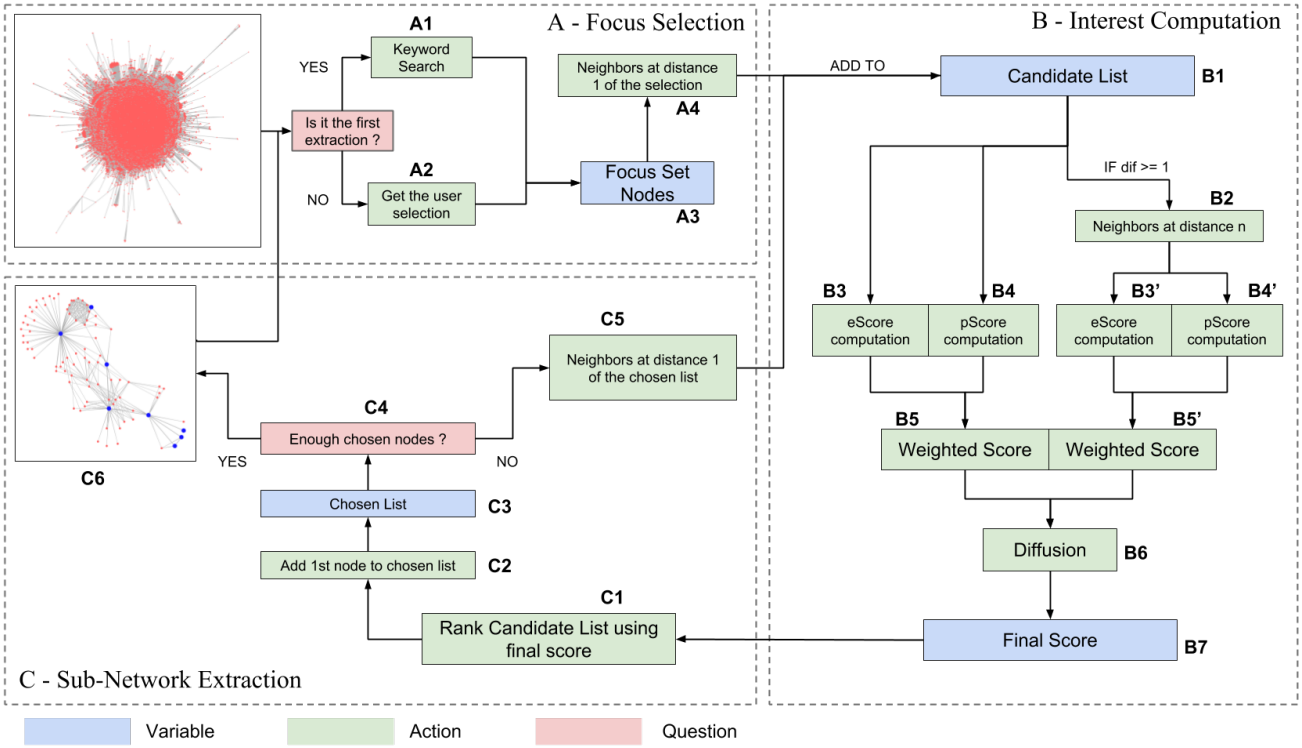


Figure 2: Pipeline of the computation of sub-networks of interest for the user from the initial 'hairball' network. This process starts with a selection operation (A) to define a candidate list i.e. a list of nodes potentially shown in the extracted sub-network. Then, a score is computed (B) to represent the interest of the user for these candidates. This score is used to rank the candidate lists. The most interesting nodes are chosen and a sub-network is build by adding every existing edges between these nodes (C). Once the sub-network is extracted and shown to the users, they (un)select one or more nodes in it and repeat all of the previous phases again until they reach satisfaction.

degree of diffusion can affect process performance depending on the data if the network is very connected. Similarly, a high percentage of diffusion equalizes all scores, making the process irrelevant. This mechanism must therefore be used with caution.

Sub-network Extraction (Fig. 2, Panel C)

This phase begins by computing the chosen list (C3). The chosen list determines which nodes are selected in the candidate list and compose the new sub-network. The nodes selected by the user are automatically in the chosen list. The objective of the whole process is therefore to fill in the chosen list according to the score in order to obtain an interesting sub-network for the user.

Through the previous phases, a score for each node was calculated. A ranking is performed (C1) and the node with the highest score is added to the chosen list (C2).

If the number of nodes in the chosen list corresponds to the number desired by the users (C4), we extract the sub-network whose nodes correspond to those of the chosen list then we add the existing links in the initial network between the extracted nodes (C6).

If not enough nodes are in the chosen list, we get the neighbouring nodes of the last node added to the chosen list and we add them to the candidate list (C5). The procedure can then be repeated from B1. A new candidate list requires scores to be calculated for unrated nodes. Nodes that have already received a score do not need to be evaluated again. The nodes are thus evaluated in the same way as for a greedy algorithm.

Once these phases have been completed, users can select new nodes in the obtained sub-network that are interesting to them. The whole process is then restarted with the previous selection enhanced with the new selection (or/and deselection) as the new candidate list (A2).

eScore: interest estimation

After having presented the overall process, we now detail and formalize the computation of the eScore (**e**xploratory **Score**, Fig. 2, Panel. B) which is computed for each iteration considering the focus set. For a given node, eScore takes into account the multilayer aspect of the network to quantify the user interest. Interest computation should differ according to the observed layer.

Formalisation

Inspired by Kivelä *et al.* [11] our model is composed of a network $G(V,L,E)$ with V the set of nodes, L the set of layers such as $\forall l \in L, l : (0,1)^{|V|}$, and E the set of edges such as $E : (V,L) \times (V,L)$. For each $v \in V$, $b_l(v) : (0,1)^{|L|}$ returns a binary vector indicating to which layers v belongs.

The user intention for each node v of V is expressed by a set F of functions. Each $f \in F$ applies on a subset of layers $L' \subseteq L$ and returns a normalized score between 0 and 1.

The eSCORE for a node x given the focus set Y is defined as

$$eScore(x|Y) = \frac{\sum_{i=1}^{|F|} f_i(x, Y, L'_i, b_l(x))}{|F|}$$

with $b_l(x) \subseteq L'_i \subseteq L$.

The role of each f function is to guide the navigation by considering semantic differences between layers or difference in the user interest between layers. We call them "*steering functions*". For instance, given a layer composed of video documents and another layer composed of audio documents. Computing a score on video document may differ from computing a score on audio document (they have different properties) and users may want audio document in priority instead of video document. By extension, for a 4-layer graph, a function assigned to the layer pattern $b_l(x) = (0,1,1,0)$ will be applicable to nodes belonging to the layer combination $(0,1,0,0)$, $(0,0,1,0)$ and $(0,1,1,0)$.

The steering functions have to be jointly defined with domain experts in order to closely match their wishes. In the next section, we propose a categorisation of these functions and we give some examples.

Steering functions categorisation

Determining an interest measure from a node can be derived either from the semantics of the data (e.g., text field search, node hand selected by the user), the topology of the network (e.g., centrality, degree, part of a clique) or, in some cases, both. Steering functions therefore follow this pattern.

In addition to this, we define two pairs of categories: the "variation category" ("focus set based" or "constant") and the "layer category" ("single layer" or "layers association based"). The steering functions are determined as a combination of each of these categories which can be also related to the topology of the network, the semantics of the data or both.

We then obtained a mathematical model of the user will for each layer or group of layers.We illustrate the following with examples from our DH data set.

Variation Type

The variation type corresponds to the level of interactivity according to the user choices made between iterations. Each steering function is either based on the user selection or static through the process.

Focus set based functions. These functions are mainly based on user selection. Because they use the focus set, their results vary during sub-network extractions according to the user's selection choices. Their goal is on the one hand to enhance interactivity by putting the user in command of the exploration process and on the other hand to allow the method to adapt if constraints on selection are to be respected during the search.

An example is the type homogeneity of a selection. Users want to have an equivalent amount of the different possible types of nodes in their selection. The selection being applied at each iteration, it is therefore necessary to propose nodes that improve the homogeneity of this selection in each new sub-network. When a resource type is dominant in the selection, the new sub-network must propose the other resource types by limiting the possible choices. This procedure is similar to an entropy optimization calculation where, by uniformising the number of each type in the selection, the sub-network shown to the user can be composed by any type of documents and entropy is then maximized. The details of the function are explained in the first scenario of the use case section.

This function is entirely based on the user selection. It minimizes the score assigned by the steering function to nodes that will probably not be selected by the user since they would degrade the homogeneity of coverage type.

Because the expert is not aware of the network structure and selects data according to his knowledge and wishes, the focus set based functions can only be semantic.

Constant functions. A constant function is a function with no user prerequisites to calculate its score. It can nevertheless be applied on one or more layers (see next paragraph). These functions can be topological or semantic.

In our network, for example, we use a rank calculation based on the degrees of all nodes (and thus on the topology of the network). A semantic example could be a proximity score calculation between a keyword and node attributes like used by Van Ham and Perer [14].

Because these calculations are independent of the context, these functions can be calculated a priori from the process M-QuBE^3 and reusable for all its iterations.

Layer Types

Layer types correspond to the application domain of the function. Each function can be either applied on a single layer or on a group of layers called layer association hereafter.

Single layer functions. Sometimes it is necessary to be able to set a specific objective for a category of nodes in the network. In our example, historians wanted to find important people linked to as many other personalities as possible in the network. So in addition to all the other functions included in M-QuBE^3, we also added an internal degree calculation to the person layer (considering only the links between two nodes belonging to the person layer).

Layer association based functions. Layer association based functions are functions allowing to highlight interactions between the different layers of the network or to make the union of certain layers around a common objective. As mentioned above, overlaps are possible between the different types of functions. So the example on entropy is also applicable for these functions. Indeed, this one applies to homogenize a set of layers, so we have the association of layer in addition to the focus set as essential parameters.

A topological example that could be used is to calculate centrality in a sub-network composed of nodes included only in a given layers association. Generally speaking, it is possible to instantiate a function that applies to an association corresponding to all the layers of the network. In doing so, classical topological functions (betweenness centrality, closeness centrality, page-rank, degree, etc.) can be used at each node of the network to compute a score.

Instantiation

The instantiation stage is therefore crucial because it determines the relevance of user guiding. For the most interactive experience possible, it is obviously preferable to favour any function using the focus set. To enhance performance, constant functions are ideal because pre-calculable and easily re-usable between iterations of the process. Finally, the single layer and layer association based functions are to be considered depending on whether our objective or our constraints apply either to a general case or to a precise part of the network.

We will illustrate more precisely the complete functioning of M-QuBE^3 in the following part by presenting more extensively our case study.

Case Study

As previously mentioned, we work in a multi-disciplinary framework with historians active in European Integration studies. In the following we detail the data and then illustrate the use of M-QuBE^3 on this data with two scenarios.

Dataset

Our historians work with a database containing documents as well as various automatically extracted information on political figures, institutions and places for a total of about 150,000 elements. Documents are heterogeneous in terms of their media types (e.g. text, image, video, audio) as well as in terms of their various forms with, for example, press releases, interviews, articles, etc. Institutions (the French state, the European Council, etc.), places (city names, country names), persons and social groups (gatherings of persons, associations) are linked to the documents that reference them for generating a network. Thus, a press article on a debate between two politicians is linked to these two politicians in the network. For media such as images or videos, if an element appears or is referenced, it will also be linked to this resource. We therefore have a graph with these different elements as nodes and the different links between these elements as edges.

All these nodes and links can be articulated in a vast multilayer network where the different types of nodes (persons, places, articles, videos, etc.) determine the layers. This data provide an ideal testing ground for M-QuBE^3: historians want to find relevant documents and entities to enrich online publications on different aspects of European integration history. Below, we detail two sample use cases using M-QuBE^3 to find new documents and elements in the database starting from a general query. These two use-cases have been suggested and validated by the historians we are collaborating with.

Steering functions definition

For optimal navigation in the network by the historians, it is necessary to understand their objectives and constraints in order to determine the steering functions.

First, historians want a homogeneous document type coverage i.e. an equivalent quantity of each type of document should be used to define a bibliography for their research. Experts interact with the M-QuBE^3 process by selecting entities they find interesting. It is therefore necessary to balance the types of documents present in the user selection. If there are too many times documents of the same type in the user selection, the homogeneity is low then it is necessary to propose document that can improve homogeneity if they are selected by the users (i.e documents of the fewest document types in the current selection). If the selection is homogeneous, the document types should be present equitably.

The objective fulfilled by the needed steering function is then to maximize the possible choices of type for the user. This problem is therefore an entropy optimization problem where the more different types selected, the more the entropy should increase. We, thus, define a function based on the Shannon's entropy applicable to the layers related to different types of documents. This function to compute a homogeneity score on n given layers is define as :

$$s(x) = 1 - \frac{\sum_{i=1}^{|Q|} Q[i]^2}{|Q| \times max_q^2}$$

with $Q : (0,V)^n$ an vector where each value of Q is the number of nodes belonging to a specific layer of n and max_q the maximal value in Q. Thus, when the number of documents of the same type is close to the maximum in the current selection, the homogeneity score $s(x)$ tends towards 1. This function then prioritizes documents that can improve document coverage to guarantee document type homogeneity to the historians.

Another element to consider is the relevance of nodes representing people and how they should be considered in the network. In their analysis work, experts want to be able to focus on the people they consider central or emblematic in the network because they play an important role in historical processes and cannot be ignored. To do this, a steering function is set on the person layer to highlight personalities highly connected to other people. This function is defined by a degree centrality but applies only to the person layer i.e. only the nodes belonging to the person layer are considered when calculating the degree of a given node. It is therefore a constant function that depends exclusively on its single application layer.

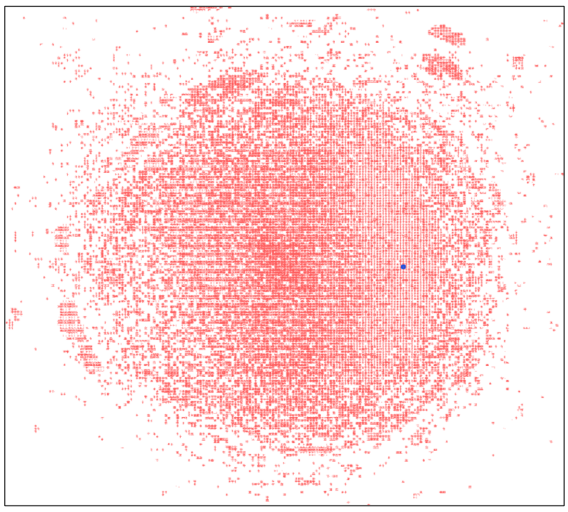
A set of other steering functions are also determined for other layers to further refine the exploration: a topological function (degree centrality) on all the layers (so the whole network) is used to give more importance to highly connected nodes in order to discover new hypothetical research paths and several constant single layer functions giving a higher importance (locations, organizations) or lower fixed importance (social groups, places) to layers on which users have no defined questions.

Once these functions are defined, historians can begin their exploration.

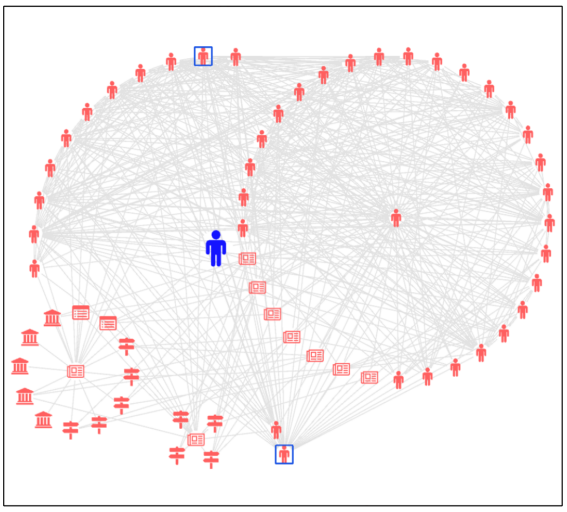
Scenario 1: Relations between Europe and USA

In this first scenario (Fig. 3), our experts work on analyzing the relations between Europe and U.S.A. (mainly economics and politics) over the years.

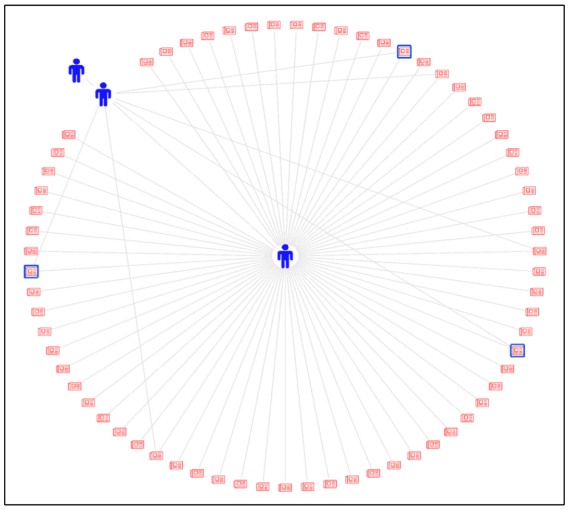
We start by selecting President George Bush by a keyword search to create a first sub-network (Fig. 3a). Only G. Bush is selected because the user does not know precisely how to approach the subject and therefore which other element to choose. George Bush's choice implicitly includes constraints. Since G. Bush belongs to a specific time and is the only element selected by the user, documents and entities that correspond to his time are privileged and have a higher probability of appearing in the sub-



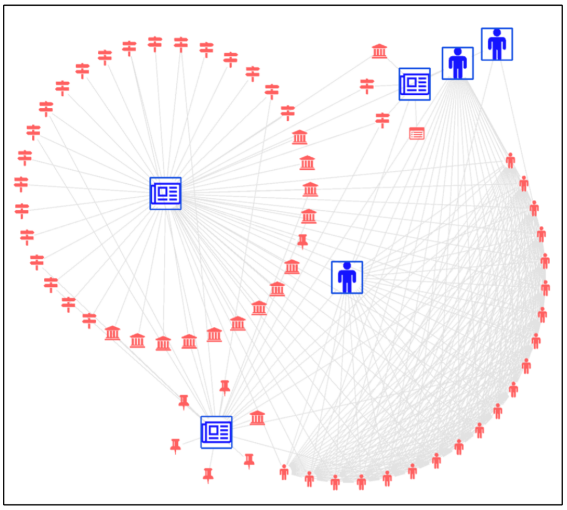
(a) George Bush is selected in the initial network by a keyword search (blue node). It is the starting point of the research process. This selection results creates the first sub-network.



(b) The user select a node representing the French presidency and a famous prime minister (square highlight). This brings new semantic information concerning both the USA (G.Bush) and the France (the new selection).



(c) This second sub-network shows a lot of documents relative to G.Bush and the French presidents (blue nodes). The user selects different documents to orient the context of his exploration.



(d) This last new sub-network proposes new tracks due to the newly selected documents. It is possible to continue on this topic or explore new paths by selecting these new elements.

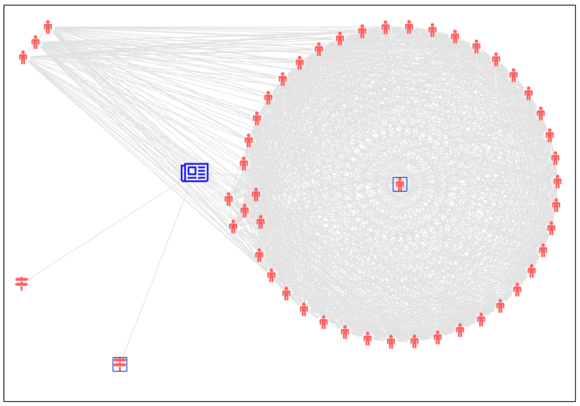
Figure 3: Relations between Europe and USA

network. The same is true of his direct political entourage as well as documents about the major events of his presidential career.

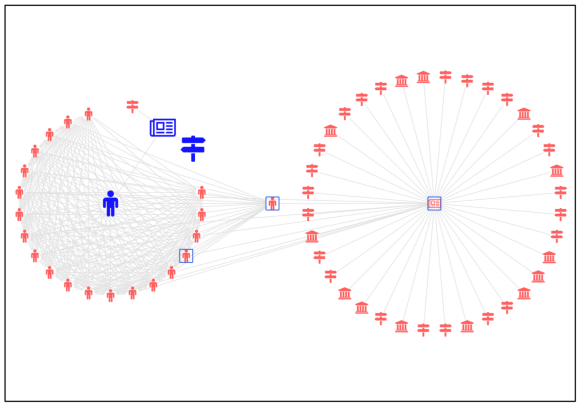
This behaviour is induced by the construction of the original network. It is progressively replaced as the user's selection is enriched by inducing a refinement of the sub-networks obtained from the new information resulting from this selection.

In this sub-network, there are a number of entities representing politicians (Fig. 3b). One of the entities represents the French Presidency. This entity aggregates relationships relating to different presidents of different eras. It acquires a high score of interest with the global and the person steering functions because it is highly connected. This node and a node representing a French prime minister are selected together in order to create the second iteration of the sub-network. Selecting the aggregated entity of the French Presidency as well as a major French politician favours a wider temporal spectrum and also makes it possible to involve the French political scene and the various related resources as potential candidates for this second iteration. This new context allows a sub-network semantically more in adequacy with the objective of the experts and thus encourages the appearance of documents that are relevant to them.

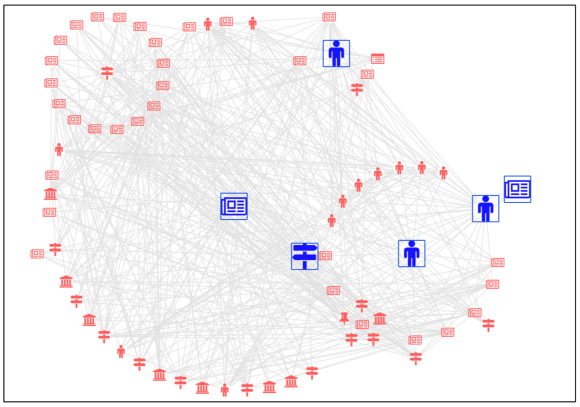
In the new extracted sub-network (Fig. 3c), new documents are shown relative to USA (G.Bush) and France (French Presidency, the Prime Minister). We select few of these documents to



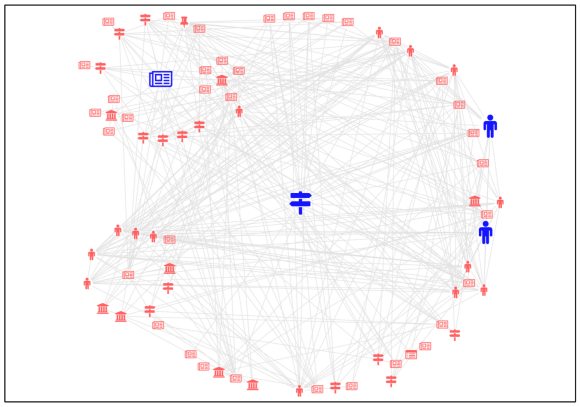
(a) We start our research with a paper about a meeting on Europe in London. We select Margaret Thatcher and London in the first sub-network (blue node).



(b) In the second sub-network, we select a document on the insertion of a new country in Europe as well as two major European actors (Jacques Delors and Pierre Werner), all linked to Margaret Thatcher (square highlight).



(c) A new path is explored by selecting the French president, the Republic of France and a document on European decadence instead of the current selection while maintaining Margaret Thatcher.



(d) In the last sub-network, the research horizon is fully renewed. It is now possible to access new documents, new places, new institutions and new persons linked directly or indirectly to the new selection.

Figure 4: The United Kingdom role in the European development

orient the context of the exploration and extract again a new sub-network.

In this third sub-network (Fig. 3d), new resources and entities appear according to the new selection. The users can then select documents related to both George Bush and French politicians that contain information about the Franco-American relationship and thus obtain new documents. They can also move to a new path and replace all selected entities with new entities. If, for example, the users are intrigued by a particular event referenced by a document "Petersberg tasks" concerning the NATO, they can replace their selection and start a new process to see the entities and potential leads that result from this event. He can, however, maintain George Bush's selection and find documents that bind them. He can also explore radically different paths by selecting the Petersburg location and discover the history of the city and why it hosted the signing of this treaty.

This iterative process is repeated until a satisfying network is found.

Scenario 2: Role of the United Kingdom in Europe's development

In this scenario (Fig. 4), our experts want to evaluate the influence of U.K. on Europe history. Our experts have as starting point a photo of a meeting held in London between Margaret Thatcher and Helmut Schmidt which is indicative of an interaction between the United Kingdom and a European Union representative.

The first sub-network is extracted only from the node corresponding to this document (Fig. 4a). In this sub-network, we select M.Thatcher and London to highlight the different political persons who were able to go to the most important place of the U.K. or interact with one of the most famous English politicians.

We get a new sub-network with a new context focused on Margaret Thatcher, London and the original document. We note the appearance of a new document about the entry of a new country into the European Union and many political figures linked to this document. Among them are Jacques Delors and Pierre Werner who are related both to Margaret Thatcher and to several other people connected to her. We select them in order to guide

the semantic context towards more information related to the major actors of Europe in order to be able to correlate them with those related to M.Thatcher (Fig. 4b).

The new sub-network generated provides a large range of new information (Fig. 4c). Documents related to the three selected actors appear and concern subjects directly related to our objective such as the 1975 English referendum or comments about M.Thatcher's vision of Europe. These documents are likely to improve type coverage thanks to our steering function and therefore mainly offer document types different from the one initially selected. Selecting these documents would insist in this direction by proposing even more entities relating to these subjects. We can also notice that these relevant documents are linked to European institutions such as the European Commission or the European Economic Community (EEC) which also appeared in this new sub-network. Selecting these entities is going to orient the documents and personalities and thus make the semantic context evolve again.

Instead, we decide to explore a new path because our curiosity drive us to evolve our initial objective. To do this, we keep from the current selection only M.Thatcher and we select a node representing the French Republic, the French President as well as a francophone article very criticizing towards Europe. The new generated sub-network (Fig. 4d) is built entirely from the new selection and then offers a new search horizon in accordance with this new path.

It is one of the specificities of the process M-QuBE^3 to propose, in addition to a great freedom of configuration and usability, to allow the experts to explore simultaneously several directions by proposing to them to change on the fly of objective. They benefit from the specification of constraints already existing with steering functions and can immediately concentrate on new emerging leads.

These examples of the M-QuBE^3 process illustrate use cases and highlight the possibility of exploring simply a dense and complex network. If no information is known on the network, it is entirely possible to set the steering functions on all the layers in a generic way with all the tools already known in graph analysis. After an initial exploration with this default configuration, it is then possible to refine the steering functions according to the new needs and thus improve the relevance of the navigation.

Discussion and Future Work

This process was built to stay as close as possible to the needs of the various experts we met. If the iterative and multilayered aspect is a first step, there are many opportunities for improvement for our programs as well as several questions that are worthy of consideration.

Although some experts are familiar with the node-link view, this view sometimes can be problematic for experts who are experiencing it occasionally or for the first time. This process is independent of its visualization and can therefore be adapted to different views adapted to the experts request. The algorithms are then applied separately on the multilayer network but the users benefit from a view more in accordance with their needs. However, it is necessary to have a selection mechanism: users must be able to select entities in each view in order to create the next view.

Another important aspect is the optimization and the diffusion. The process acts like a greedy algorithm. However, if a high degree of diffusion is used, or if the network is very connected, the whole or a large part of the network may need to be evaluated. Depending on the size of the network, a significant drop of performance can be noticed. The limitations of this diffusion have not been yet evaluated. However, this diffusion is only used to avoid the isolated relevant node situation (where an interesting node cannot be reached because its neighbors do not have a sufficient score to be selected). It is therefore quite possible to use a zero degree diffusion if such a scenario is not a problem for experts.

Finally, one of the opportunities offered by this process is the exploitation of emerging leads. While navigating, you can change your selection or return to a previous state of the selection to try other paths. For that purpose, we use in our program an additional network where each node is a state of the selection and a simple click on these nodes allows you to return to this state of the selection and see the associated sub-network. The experts are very enthusiastic about this feature and different ideas are currently explored to improve its use. This would bring the M-QuBE^3 process even closer to the needs of our experts and their specific methodologies.

Finally, while we have positive feedback and scenarios validated by experts, we have not yet established formal evaluation procedures. This will be useful to experience further the method and find new opportunities for improvement.

Conclusion

We have presented in this paper a multilayer graph exploration method for data experts. Through an application, implemented through a thick client implementation and a lighter online version, experts can easily navigate a network without having to learn its structure or content. They are directly in charge of the research by selecting the steering functions according to their constraints or their wishes and by progressively specifying, as subnetworks are extracted, the desired context for their navigation and their objectives. This way of proceeding makes it possible to put forward a flexibility which is often necessary for the human sciences or for blind exploration. In addition, the user explores the network with a partial view, made to be representative of what he wants to see and easily analysable, allowing a qualitative study of the network without excluding the possibility of following new leads. In conclusion, M-QuBE^3 offers an effective way for exploring multilayer networks that would be too complex for traditional methods while at the same time offering experts a way of delving into data that fits with the iterative operating method specific to many domains.

Acknowledgements

This work was (partially) funded by the ANR grant BLIZAAR ANR-15-CE23-0002-01 and the FNR grant BLIZAAR INTER/ANR/14/9909176. Thanks to historians of the CVCE and particularly to Marten DURING for the constructive working sessions as well as Fintan Mc Gee and Mickaël Stefas for their implementation work and their support.

References

[1] Beel, J., Gipp, B.: Google scholar's ranking algorithm: an introductory overview. In: Int. Conf. on Scientometrics and Informetrics (ISSI'09). vol. 1, pp. 230–241 (2009)

[2] Card, S.K., Nation, D.: Degree-of-interest trees: A component of an attention-reactive user interface. In: AVI '02. pp. 231–245. ACM (2002)
[3] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al.: The youtube video recommendation system. In: Proc. of the 4^{th} ACM Conf. on Recommender systems. pp. 293–296. ACM (2010)
[4] De Domenico, M., Porter, M.A., Arenas, A.: Muxviz: a tool for multilayer analysis and visualization of networks. Journal of Complex Networks **3**(2), 159–176 (2015). https://doi.org/10.1093/comnet/cnu038
[5] Furnas, G.W.: Generalized fisheye views. In: SIGCHI Conf. '86. pp. 16–23. ACM (1986)
[6] Ghoniem, M., Fekete, J.D., Castagliola, P.: A comparison of the readability of graphs using node-link and matrix-based representations. In: Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on. pp. 17–24. Ieee (2004)
[7] Ghoniem, M., Fekete, J.D., Castagliola, P.: On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. Information Visualization (Palgrave) **4**(2), 114–135 (2005)
[8] van Ham, F., van Wijk, J.J.: Interactive visualization of small world graphs. In: IEEE Symp. Infor. Vis. pp. 199–206 (2004)
[9] Hüsken, P., Ziegler, J.: Degree-of-interest visualization for ontology exploration. Human-Computer Interaction–INTERACT 2007 pp. 116–119 (2007)
[10] Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. In: Information visualization, pp. 154–175. Springer (2008)
[11] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. J. of Complex Networks (2), 203–271 (2014). https://doi.org/10.1093/comnet/cnu016
[12] McGee, F., During, M., Ghoniem, M.: Towards visual analytics of multilayer graphs for digital cultural heritage. Towards Visual Analytics of Multilayer Graphs for Digital Cultural Heritage (2016)
[13] Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: IEEE Symp. on Vis. Lang. pp. 336–343 (1996)
[14] Van Ham, F., Perer, A.: "search, show context, expand on demand": Supporting large graph exploration with degree-of-interest. IEEE Trans. on Vis. and Comp. Graph. **15**(6), 953–960 (2009)

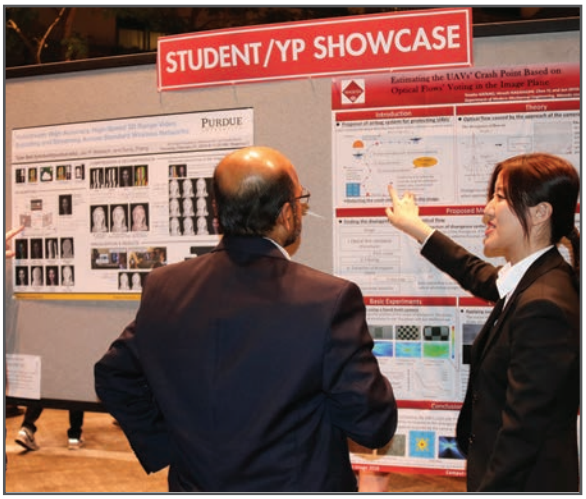
JOIN US AT THE NEXT EI!

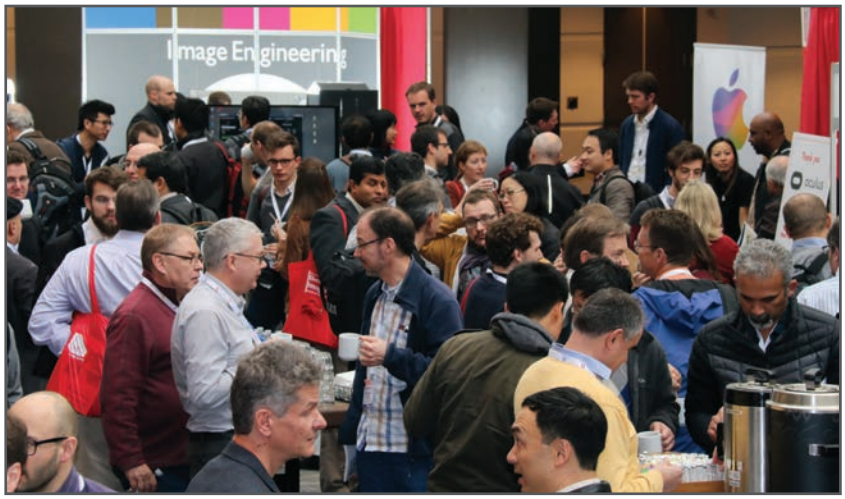
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org