Mining Networks through Visual Analytics: Incremental Hypothesis Building and Validation

David Auber a, Romain Bourqui a and Guy Melançon a,1

a INRIA Bordeaux Sud-Ouest / CNRS UMR 5800 LaBRI
Campus Université Bordeaux I
351 Cours de la Libération
33405 Talence Cedex – FRANCE

Abstract. Interactive visualization supports the analytical process: interacting with abstract views, using the data as a malleable material, analysts build hypothesis that can be further validated on the whole dataset. We use graph clustering in order to group elements and show them as meta-nodes and reduce the number of visual items, further organizing data over into several layers, in an effort to provide a multi-level model of the studied phenomenon. In doing so, hierarchical clustering not only contributes to the study of data but also brings in ingredients for interaction, enabling the user to zoom in and out of clusters while exploring the data in quest for evidence.

Keywords. Visual Analytics, Graph Mining, Information Visualization, Interactive Exploration

Introduction

Networks are everywhere. Social networks: people know each other, people meet, people exchange information. People rely on third parties belonging to different organizations, people link organizations. Events first group into sequences and then into causal networks when people collaborate. Semantic networks: documents link because they share content or authors, or through citations. Concepts and ideas link through documents - people create, access, modify and share documents. Analysts are faced with massive collections gathering documents, events and actors from which they try to make sense. That is, they search data to locate patterns and discover evidence. Interactive exploration of data has now established as a fruitful strategy to tackle the problem posed by this abundance of information. The Visual Analytics initiative promotes the use of Information Visualization to support analytical reasoning through a sense-making loop based on which the analysis incrementally builds hypotheses (Fig. 1).

"Information Visualization" supports the visual exploration and analysis of large datasets by developing techniques and tools exploiting human visual capabilities – ac-

1Corresponding Author: INRIA Bordeaux Sud-Ouest / CNRS UMR 5800 LaBRI; E-mail: Guy.Melancon@labri.fr
40% of our cortical activities are dedicated to processing visual signals. The design of new visualization methods and tools becomes even more necessary with the continuously increasing volume of available data, which poses a problem that obviously cannot be solved by relying solely on the increase of CPU power.

Visually mining data requires combining data analysis with visual graphics and interaction. Mining itself draws not only on statistics but in a rather astute mixture of mathematical rigor and heuristic procedures. As David Hand puts it [12] [13]:

“To many, the essence of data mining is the possibility of serendipitous discovery of unsuspected but valuable information. This means the process is essentially exploratory.”

From Hand’s perspective, we see that information visualization has much to share with data mining because visualization often comes as an aid to exploratory analysis. The perspective we adopt is a combination of (semi) automated data processing together with human analytical and perceptual capabilities. Although relying on technology, the analysis task remains in total control of the human user. The National Visual Analytics Center (NVAC) research agenda [16] clearly states:

“[The] analysis process requires human judgment to make the best possible evaluation of incomplete, inconsistent, and potentially deceptive information […]”,

later calling for the development of

“[…] visually based methods to support the entire analytic reasoning process . . . .”

That is, in ideal cases the visualization should be designed in order not only to assist the analysis but to also actively contribute to its progress. Visual Analytics thus appears as a multi-disciplinary field embracing a large spectrum of competences. This partly comes from the need to cover all processes involved in the so-called “Visualization pipeline” as depicted by dos Santos and Brodlie [7] (Fig. 2):
1. Interaction, Scalability, Graph Hierarchies

Visual Analytics obeys Daniel Keim’s mantra "Analyse first - Show the Important - Zoom, Filter and Analyse Further - Details on Demand". Showing the important can be understood in several different manners; what remains essential here is to enable the user to dynamically build views from the original dataset under study. For example, the graph shown in Fig. 3 has been extracted from a NCTC dataset (consisting of 10 000 nodes and 20 000 edges) to help locate collaboration between terrorists groups based on territorial activity. Continents have been inserted into the graph to help organize the overall layout. Drawing the whole graph with tens of thousands of elements on the screen is pointless, as the resulting drawing obviously lacks readability. Once this smaller graph has been drawn and explored, a second iteration can be designed based on a different point of view, building on previous observations. The overall strategy is thus to filter out data elements and build sequences of "virtual" or "partial" graphs, in an attempt to "see" what can possibly be present in the data.

The visualization supports the analytical process: ultimately, the analyst will build hypothesis that can be further validated on the whole dataset. It is important to note that this incremental methodology uses the data as a malleable material. The analysis of "virtual" graphs offering partial views on the dataset can be useful in bringing structural properties upfront. Indeed, having established that the graph under study is small world, or scale-free, can trigger various scenarios to further analyze the data.

1.1. Hierarchical clustering

The number of data elements drawn on the screen must necessarily be kept small, for sake of readability. A layout of a graph containing a thousand nodes is already hard to read, depending on its structural properties (a tree is much easier to read than a totally random graph, of course). A common strategy is then to use graph clustering in order to group elements and show them as meta-nodes and reduce the number of visual items. Promising approaches develop techniques based on the intuitive notion of intracluster density versus intercluster sparsity. Methods based on the computations of node or edge centralities [8] [3] have proved successful [14] [10]; other approaches based on local indices (edge strength [2] [5] or Burt’s constraint on edges and nodes [4], for instance) have also been suggested.

Typically, hierarchical clustering will organize data over several layers. In ideal cases, these layers will contribute to model the data under study as layers reveal structure. Graph hierarchies appear as an adequate formalization capturing the notion of multiscale communities in a network (network of networks). Current studies confirm the absolute presence of hierarchies either in nature itself or in abstract human construction such as language [9]. Current evolutionary models in biology try to capture the multilevel nature.
Figure 3. The drawing of a graph sketching the relationships among terrorist groups and countries. Continents are further inserted to organize the layout.

of networks formed by various biological entities [18] [15], just as with cities and city systems in geography.

Also, hierarchical clustering not only contributes to the study of data but also brings in ingredients for interaction. Our methodology for studying large small world networks relies on graph hierarchies (nested subgraphs) as a central paradigm. For example, Fig. 4 below displays a hierarchy of subgraphs computed out of a network of keywords extracted from documents (keywords are linked based on a similarity measure; a graph is then induced by thresholding the similarity matrix). As can be seen on the figure, the top level graph consists of meta-nodes themselves containing a lower level hierarchy of graphs. This visual representation can then be zoomed in and out in order to explore more focused regions of the graph. The top level graph shows the overall organization of the network. Meta-nodes are connected to one another according to connections between the nodes they contain.

Once a graph hierarchy has been computed from the original dataset, it can be used to compute various statistics on data elements - either to assess of the relevance of the hierarchy or simply to explore properties of low level nodes. We recently revisited the work by Guimera et al. [11] extending the z-score and participation coefficient to weighted graphs and to graph hierarchies. The z-score corroborates an individuals’ dynamic within its own community, while its participation coefficient indicates how much it covers all
other communities. When dealing with multiscale networks, the computation of the z-score and participation coefficient of individuals and communities, at various levels, reveals how the network’s dynamic build through scales. Appropriate visual cues help locate key actors, pointing at individuals or communities as hubs, bridges or satellite.

1.2. Tulip

The Graph Visualization Framework Tulip\(^2\) [1] [6] developed by our team includes graph hierarchies as a central navigation mechanism and data structure. Variants of hierarchical clustering algorithm or graph statistics can readily be implemented as plug-ins and used on the spot. The overall architecture exploits all capabilities of graphics hardware and C++, making it one of the most powerful publicly available graph visualization framework.

References


\(^2\)Tulip is Open Source and distributed under GPL. See the URL www.tulip-software.org