Knowledge Maps of Web Graphs

Valeria Fionda[1], **Claudio Gutierrez**[2], **Giuseppe Pirrò**[3]
[1] Department of Mathematics, University of Calabria
[2] DCC, Universidad de Chile
[3] KRDB, Free University of Bolzano

Abstract

In this short note we give an overview of our research concerning cartography on the Web and its challenges. We present a mathematical formalism to capture the notion of map on the Web, which allows to automatize the construction of maps.

1 Introduction

Cartography is the art of making maps. Its essential aim is that of representing the characteristics of a region of interest. Cartography builds upon two main steps: *selection* and *abstraction* (Robinson et al. 1995). Selection enables to focus only on the particular pieces of information that will serve the map's purpose; in this phase the region of the space to be mapped is chosen. Abstraction is the fundamental property of a map, which states that a map should be smaller than the region it portraits. Thus, a map can be simply defined as an abstract representation of a region of interest.

The Web is a large and interconnected information space commonly accessed and explored via navigation. This space is simply too large and its interrelations too complex for anyone to grasp its content by direct observation. Hence, the possibility of applying cartographic principles to the Web space becomes a relevant matter. Knowledge maps can be useful cues that help to navigate, find routes and discover previously unknown connections between knowledge items on the Web. Effectively, they can play the role of *navigational charts* helping users to cope with the size of the Web (cyber)space (Dodge and Kitchin 2001). Users via knowledge maps can track, record, identify and abstract conceptual regions of information on the Web, for their own use, for sharing/exchanging with other users and/or for further processing (e.g., combination with other maps). Maps are also useful to analyze information. For instance, the availability of a series of chronologically sequential maps enables complex map analysis (e.g., longitudinal analysis) for the detection and forecasting of trends in specific domains (Garfield 1994). This is useful, for instance, in the analysis of the knowledge flows in scientific literature, which helps in understating how the interlinking between disciplines is changing (Rosvall and Bergstrom 2008). Maps of social networks

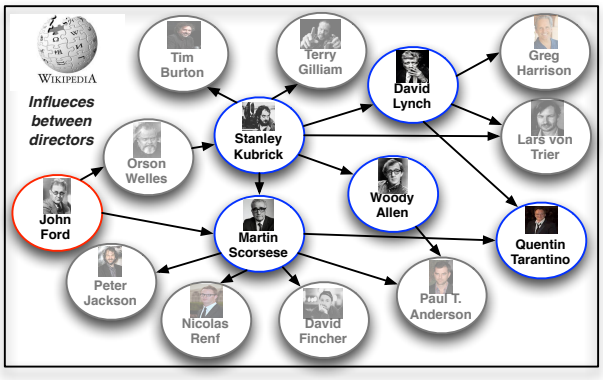

Figure 1: A region of the Web taken from Wikipedia.

can be analyzed to forecast friendship (Ledbetter, Griffin, and Sparks 2007).

The progresses in Web technologies and languages originating from the Semantic Web proposal and, in particular the availability of structured and machine-readable information (RDF data model and SPARQL language) open new perspectives toward the feasibility of knowledge maps of the Web. Now it is possible to leverage logical/algebraic methods to organize and specify in a formal, machine-understandable and reusable way regions and maps of the Web. In this short note we introduce the foundations of a formal notion of map for the Web that will enable the interpretation of maps not only by humans but also by machines, fostering in this way their exchange, combination and reuse at Web level. The following example gives an idea of how a knowledge map of the Web looks like.

Example 1 *(**Mapping favorite directors**). Syd is fond of cinema and wants to discover and keep track of his favorite directors and relations among them.*

The first step toward building the map of *Syd*'s favorite directors is to *select* the region of interest. This could be done by *tracking* the navigational activity of *Syd* on Wikipedia; Syd starting from the page of one of his favorite directors (i.e., J. Ford) has navigated toward directors influenced by J. Ford and manually (book)marked some of them, that is, S. Kubrick, W. Allen, D. Lynch, M. Scorsese and Q. Tarantino.

Note that the region besides these *distinguished* nodes also contains other (lighter) nodes that were visited during the navigation (see Fig. 1). Once the region in Fig. 1 has been obtained, it has to be *abstracted* in a knowledge map. Fig. 2 shows two possible maps of the region in Fig. 1. Which one is a "better" representation of the original region?

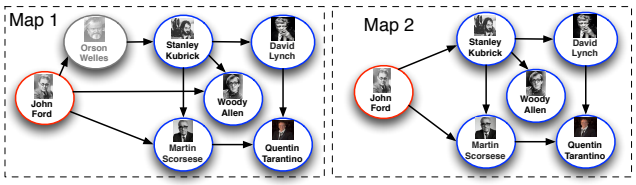

Figure 2: Two possible maps of the region in Fig. 1.

The challenge is to identify a suitable definition of knowledge map that provides an economic and informative representation of the region of interest. Essentially, a map must reflect in a *concise way* information in the region in terms of *connectivity* among some distinguished nodes. What are the formal property of knowledge maps of the Web? Answering these questions enables to have a clear understanding of the relations between regions and maps and also to define means to represent, exchange, reuse and combine different maps (i.e., via an algebra). In this short note we provide a formalization of the notions of region and map of the Web. We discuss several types of maps, present algorithms for constructing such maps and a concrete example of Web map.

2 Formalizing Maps

One of the main challenges toward building maps of the Web is how to move from human-readable maps towards maps that can also be understood and manipulated by machines. This requires the development of a formal notion of map and the investigation of its properties/relations with the region it represents. Let us introduce some notations and basic definitions. Let $\Gamma = (V, E)$ be a graph where V is the set of nodes and E is the set of edges. Then:

- $u \to v$ denotes an edge $(u, v) \in E$. As an example, S. Kubrick $\to$ W. Allen in Fig. 1.
- $u \twoheadrightarrow v$ denotes a path from u to v in Γ. As an example, J. Ford $\twoheadrightarrow$ S. Kubrick in Fig. 1.
- Let $N \subseteq V$. Then, $u \twoheadrightarrow_N v$ if and only if there is a path from u to v in Γ not passing through intermediate nodes in N. As an example, if $N = \{$M. Scorsese, W. Allen$\}$ then J. Ford $\twoheadrightarrow_N$ S. Kubrick but not J. Ford $\twoheadrightarrow_N$ P. T. Anderson in Fig. 1.

Definition 2 (Map) *Let $\Gamma = (V, E)$ be a graph. A* map *$M = (V_M, E_M)$ is a graph s.t. $V_M \subseteq V$ and each edge $(x, y) \in E_M$ implies $x \twoheadrightarrow y$ in Γ.*

By making a parallel with cartography, the graph Γ is the region (or "territory") to be represented/abstracted by the map M. The set of nodes V_M in M represent "points" relevant for the map; these are the *distinctive traits* used as the basis to *identity of the region*. The set of edges E_M in the map M represent the "directions", "routes" and signals connecting different points of interest. A basic (and highly used) example of maps of the Web are bookmarks. In this case, V_M is the set of nodes highlighted or marked, and $E_M = \emptyset$, that is, there is no connectivity recorded among them. More elaborate maps can be built from citation graphs (e.g., Google Scholar); here V_M is a set of URIs of papers of interest and E_M is the citation connectivity graph one can navigate.

Fig. 3 shows some possible maps obtained from the graph in Fig. 1. The map in Fig. 3 (a) is an example of bookmarks. Note that maps (a)-(c) do not capture connectivity among the distinguished nodes. For instance, J. Ford and S. Kubrick that are both distinguished nodes connected by a path in the original region, are not connected in these maps. This points to the notion of completeness of a map.

Definition 3 (Complete Map) *A* map *is complete if $\forall x, y \in V_M$, $x \twoheadrightarrow y$ in Γ implies $x \twoheadrightarrow y$ in M.*

In Fig. 3, maps (d)-(f) are examples of complete maps. However, even completeness is not enough to abstract information. Consider the map in Fig. 3 (d). In the region in Fig. 1 there is an edge connecting J. Ford and M. Scorsese. In the map, although there is a path connecting (via S. Kubrick) the two nodes, the fact that there is an alternative connection (not passing through nodes in V_M) between them is missing. The following notion captures this issue:

Definition 4 (Route-complete map) *A* map *M of Γ is route-complete iff $\forall x, y \in V_M$ $x \twoheadrightarrow_{V_M} y$ in Γ implies $x \to y$ in M.*

By looking at Fig. 3, map (d) is not route-complete while maps (e) and (f) are route-complete. Indeed, for each path between distinguished nodes not passing through any other distinguished node in the region, there is an edge in the map. As an example, maps (e) and (f), both introduce some direct edges like, for instance, J. Ford$\to$S. Kubrick (i.e., e_1) that were not present in the original region; the edge e_1 *abstracts* the path J. Ford$\to$O. Welles$\to$S. Kubrick.

But there are still some issues to address. Consider the edge e_2 between J.Ford and Q.Tarantino in map (f). All paths connecting J.Ford to Q.Tarantino in the region pass through some distinguished nodes (i.e., M.Scorsese, S.Kubrick, D.Lynch) and connectivity information between J.Ford and Q.Tarantino is already encoded in the map via the paths Ford $\to$ Scorsese $\to$ Tarantino, Ford $\to$ Kubrick $\to$ Scorsese $\to$ Tarantino and Ford $\to$ Kubrick $\to$ Lynch $\to$ Tarantino.
Thus, edge e_2 does not add new connectivity information, i.e., it is in some sense redundant.

Definition 5 (Non-redundant map) *A* map *M of Γ is non-redundant iff $\forall x, y \in V_M$ $x \to y$ in M implies $x \twoheadrightarrow_{V_M} y$ in Γ.*

Example of non-redundant maps are reported in Fig. 3 (d) and (e). However, recall that (d) does not satisfy the property of route-completeness. In order to capture completeness, route-completeness and non-redundancy we introduce *good maps*.

Definition 6 (Good map) *A* map *M of Γ is good if and only if it is complete, route-complete and non-redundant.*

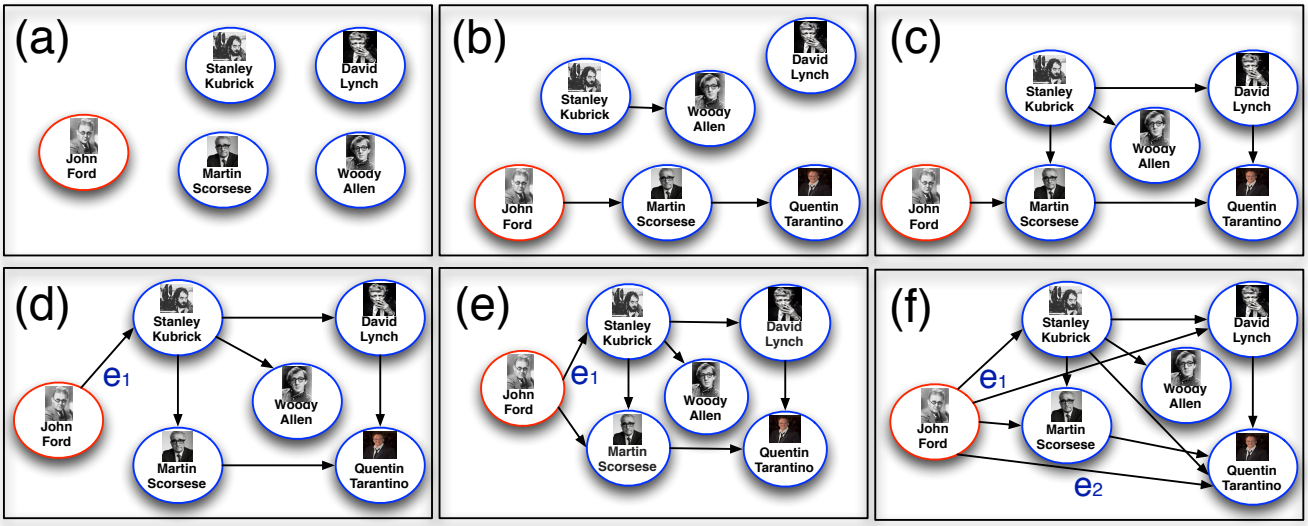

Figure 3: Examples of maps obtained from the region and the distinguished nodes (the set V_M) in Fig. 1.

Thus, good maps comprise the basic features we would like to have in a map in order to faithfully and economically capture connectivity. Fig. 3 (e) shows the good map of the region in Fig. 1. Indeed it is the *unique* good map of such a region. This is not a coincidence. Next theorem give a simple algorithmic characterization of good maps and shows that they are unique:

Theorem 7 *Let $\Gamma = (V_\Gamma, E_\Gamma)$ be a region. Then:*

1. A map M of Γ is good iff $\forall x, y \in V_M$, $x \to y$ in $M \Leftrightarrow x \twoheadrightarrow_{V_M} y$ in Γ.

2. Given $V_M \subseteq V_\Gamma$, there is a unique good map $M = (V_M, E_M)$ over Γ.

Good maps behave very well from algebraic and computationally point of views. In this short note we do not have space to exhibit further formal properties (operations between them, algorithms for computing, etc.), which will be presented in an extended version.

3 Maps for the Web of Data

Our motivation behind the formalism for maps is to develop knowledge maps of the Web graph. The availability of structured interlinked data on the Web (in the form of RDF) is evolving the traditional Web of Documents into a semantic distributed graph known as Web of Linked Data (WoD) (Heath and Bizer 2011), which we will denote as $\mathcal{W} = (V, E)$. Here, nodes are RDF data sources (identified by URIs) and edges RDF predicates. In this section we discuss how to instantiate the conceptual framework introduced in the previous section in the Web of Linked Data.

The problem of building maps in this scenario, given the machinery to build maps already presented, can be reduced to identify regions of interest:

Problem 1 *Given the WoD graph $\mathcal{W} = (V, E)$, construct a subgraph (region) $S = (V', E')$ of $\mathcal{W}$ that contains some distinguished nodes of interest $N \subseteq V'$.*

Graph navigational languages partially address this problem. Current navigational languages (e.g. XPath (W3C 1999), nSPARQL (Pérez, Arenas, and Gutierrez 2010), NautiLOD (Fionda, Gutierrez, and Pirrò 2012), etc.) enable to navigate a labeled graph essentially via matching some pattern (or navigational expression) expressed by means of a regular expression over the alphabet of edge labels. Their output is a set of nodes satisfying the matchings. For our purposes here, their drawback is that the information used during the navigation is forgotten, that is, the semantics uses the navigation only as the means to reach the resulting *set of nodes*. This is not enough for our goal: we need the missing information about connectivity in order to deal with regions and maps.

To cope with this problem, we extend the semantics of navigational languages to incorporate in their results not only the final set of nodes, but some of the paths followed during the evaluation. This allows to build regions instead of getting only sets of nodes. We consider the navigational language $\mathcal{L}$ that abstracts the core of most existing path languages. Its syntax is:

`path::=<RDF predicate>| path[test] | (path)* | (path|path) | path/path`

The core of $\mathcal{L}$ are regular expressions built upon RDF predicates and boolean tests used to drive the navigation. Tests can be of different types, for instance, ASK SPARQL queries (as in NautiLOD (Fionda, Gutierrez, and Pirrò 2012)) and/or nested regular expressions (Pérez, Arenas, and Gutierrez 2010). We are now ready to provide a concrete example of map.

Example 8 *Specify a region that contain scientists that have been influenced, directly or indirectly up to distance 6, by Tim Berners-Lee (TBL). Build the good map of the obtained region.*

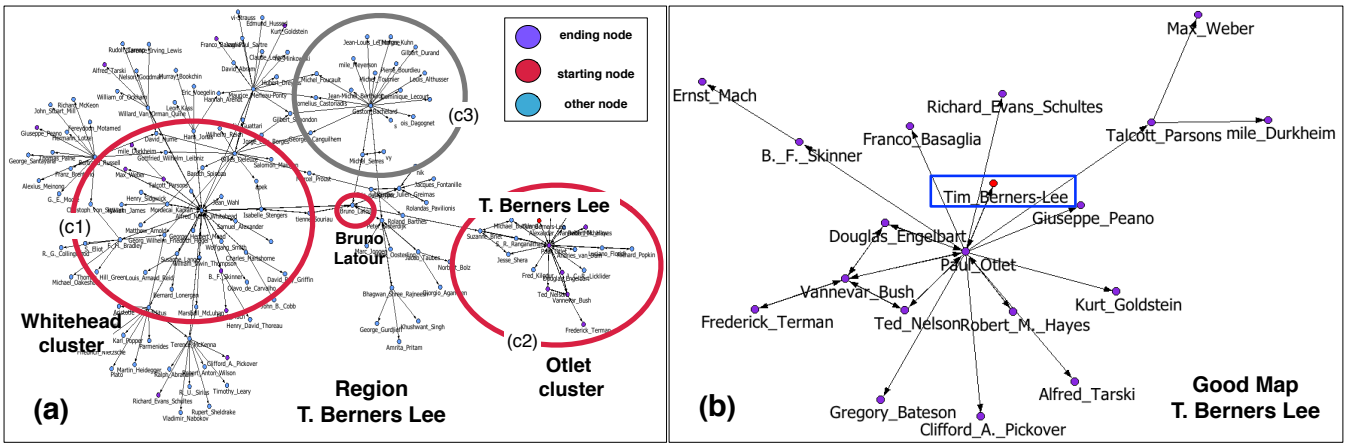

Figure 4: Region (a) and good map (b) for Example 8.

The region can be specified by considering the URI of TBL in DBpedia as starting node and the following $\mathcal{L}$ expression:

```
dbp:influenced<1-6>[test]
```

`test=` `ASK {?person rdf:type dbpedia:Scientist.}`

The notation `<1-6>` is a shorthand for the concatenation up to six steps of the predicate `dbp:influenced`, while the `test` is used to filter the endpoint nodes representing scientists. Fig. 4 (a) reports the region which contains 149 nodes and 236 edges. There are two main influence clusters; the first (c2) around Otlet and the second one (c1) around Whitehead. Note also that Latour is the bridge between the two clusters but (according to `dbpedia.org`) he is not a scientist. In the region there are several nodes that were visited during the evaluation of the expression but did not lead to any result (e.g., the cluster of nodes c3). Fig. 4 (b) shows the good map. It has 18 nodes and 43 edges and represents the connectivity information among scientists in the region.

4 Related Work

The development of our map framework share some characteristics with other approaches that focus on providing some form of "map" of the Web. Dodge (Dodge and Kitchin 2001) in the *Atlas of the Cyberspace*, provides a comprehensive overview of visual representations of *digital landscapes* on the Web. A recent information visualization paradigm used to summarize information is that of *metro maps* (e.g., (Shahaf, Guestrin, and Horvitz 2012)). Other strands of research related to ours are (visual) navigational histories, site maps and bookmarks. An approach enabling to create concept maps is described in Gaines and Shaw (Gaines and Shaw 1995). As for site maps, Li et al. (Li et al. 2001) investigated the problem of building multi-granular maps. Finally, bookmarking consists in marking and sharing (e.g., with Delicious) URIs for the purpose of future reuse.

The crucial difference with the present work is that these approaches are designed for human usage and are mainly oriented to visualization. Moreover, no formal/provable relations of connectivity between the URIs chosen are given.

Acknowledgments Fionda has been supported from the European Commission, European Social Fund and the Calabria region. Gutierrez was supported by the grant FONDECYT No 1110287.

References

DeRose, S. and Clark, J (eds). 1999. XML Path Language (XPath) W3C Recommendation.

Dodge, M., and Kitchin, R. 2001. *Atlas of Cyberspace*. Addison-Wesley Great Britian.

Fionda, V., Gutierrez, C., and Pirrò, G. 2012. Semantic Navigation on the Web of Data: Specification of Routes, Web Fragments and Actions. In *Proceedings of the 21st World Wide Web Conference (WWW)*, 281–290. ACM.

Gaines, B., and Shaw, M. 1995. WebMap: Concept Mapping on the Web. *World Wide Web Journal, 1(1), 171-183* 33.

Garfield, E. 1994. Scientography: Mapping the Tracks of Science. *Current Contents: Social & Behavioural Sciences* 7(45):5–10.

Heath, T., and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool.

Ledbetter, A. M.; Griffin, E.; and Sparks, G. G. 2007. Forecasting "friends forever": A longitudinal investigation of sustained closeness between best friends. *Personal Relationships* 14(2):343–350.

Li, W.; Ayan, N.; Kolak, O.; Vu, Q.; Takano, H.; and Shimamura, H. 2001. Constructing Multi-Granular and Topic-Focused Web Site Maps. In *WWW*, 343–354. ACM.

Pérez, J.; Arenas, M.; and Gutierrez, C. 2010. nSPARQL: A Navigational Language for RDF. *Journal of Web Semantics* 8(4):255–270.

Robinson, A. H.; Morrison, J.; Muehrcke, O. C.; Kimerling, A.; and Guptill, S. C. 1995. *Elements of Cartography*. Wiley.

Rosvall, M., and Bergstrom, C. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105(4):1118–1123.

Shahaf, D.; Guestrin, C.; and Horvitz, E. 2012. Trains of Thought: Generating Information Maps. In *Proceedings of the 21st World Wide Web Conference (WWW)*, 899–908. ACM.