Building Knowledge Maps of Web Graphs

Extended Abstract*

Valeria Fionda
DeMaCS, University of Calabria
Italy
fionda@mat.unical.it

Giuseppe Pirró
ICAR-CNR
Italy
pirro@icar.cnr.it

Claudio Gutierrez
DCC, Universidad de Chile & CIWS
Chile
cgutierr@dcc.uchile.cl

ABSTRACT

We research the problem of building knowledge maps of graph-like information. We live in the digital era and similarly to the Earth, the Web is simply too large and its interrelations too complex for anyone to grasp much of it through direct observation. Thus, the problem of applying cartographic principles also to digital landscapes is intriguing. We introduce a mathematical formalism that captures the general notion of map of a graph and enables its development and manipulation in a semi-automated way. We describe an implementation of our formalism on the Web of Linked Data graph and discuss algorithms that efficiently generate and combine (via an algebra) regions and maps. Finally, we discuss examples of knowledge maps built with a tool implementing our framework.

ACM Reference Format:
Valeria Fionda, Giuseppe Pirró, and Claudio Gutierrez. 2018. Building Knowledge Maps of Web Graphs : Extended Abstract. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3186237

1 INTRODUCTION

The Web is a vast space of interconnected information that users commonly access via navigation enabled by browsers. However, the Web is simply too large and its interrelations too complex for anyone to grasp much only by direct observation. Consider the task of navigating a citation network by using, for instance, Google Scholar. One typically starts from a seed paper. Then, by clicking on the cited by link, s/he navigates towards papers that have cited the seed paper, selects those of interest (e.g., by bookmarking them) and continues. After a while it is very hard to reconstruct the network of citations in terms of papers of interest and connections between them. Moreover, the whole process is manual. Having an automatic way of identifying the portion of the citation network of interest (i.e., papers and their connections) and some form of abstract representation, where only salient papers (e.g., papers with certain keywords in the title) and links between them are represented, would be an extremely useful support to the navigation.

To cope with the huge amount of interconnected information available on the Web, we take inspiration from cartography and introduce a framework to build maps of the Web. In the physical space, the process of map making can be summarized in two main steps, that are selection and abstraction [10]. *Selection* enables one to focus only on the particular pieces of information that will serve the map's purpose; specifically, in this phase the region to be mapped is chosen. In our previous example about navigating a citation network, the region would consist in nodes (i.e., papers) and cited by links visited during the navigation. *Abstraction* is the fundamental property of a map, which states that a map is always smaller than the region it portrays. Abstracting our citation network could be done by considering only nodes with certain properties (e.g., papers published in some specific conference) and links between them.

Recent progress in Web technologies and languages originating from the Semantic Web proposal as well as the availability at planetary scale of structured information in the Resource Description Framework (RDF) standard data format, open new opportunities toward automating the construction of Web Maps. On one hand, interlinks between data items, encoded in RDF predicates, carry a precise semantic meaning, thus allowing for precise characterization of the nature of reachability that is crucial in extracting Web regions. On the other hand, maps can be given an RDF representation and then be processed not only by humans, via visual interpretations, but also by machines, due to the machine-processable nature of RDF. This will foster the exchange, combination and reuse of maps. We believe that the availability of Web Maps can help users coping with the complexity of the Web regions in the same way as geographic maps help users cope with the complexity of large physical regions.

There are many tools (partially) touching these aspects. The most traditional and popular are bookmarks: a list of URLs, sometimes categorized by tags. This idea has been enhanced to incorporate, for instance, social features (share, rank, tag bookmarks) and/or annotations of different types of data (e.g., not only URLs but also documents). Delicious and Diigo are two popular bookmarking systems. Other approaches go beyond bookmarks and enable to organize URLs to also highlight connections between them. Results are grouped and presented in the form of a graphical map. Some examples are search engines like Tag Galaxy, navigational history tools (e.g., [3]), visual HTML site maps (for users) and atlases of the Web (e.g., [2]). More recent approaches focus on providing visual representations of specific domains such as publications or news (e.g., [11]). Existing approaches do not comply with the idea of a map that we envision. First, they do not build maps in the cartographic sense as they miss the abstraction phase. Second, they are designed for human visualization only. Third, they do not enable the declarative specification of the region (e.g., portion of interest of the Web) to be mapped. Fourth, they lack a formal mathematical model.

*An extended version appeared in Artificial Intelligence [7]

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186237

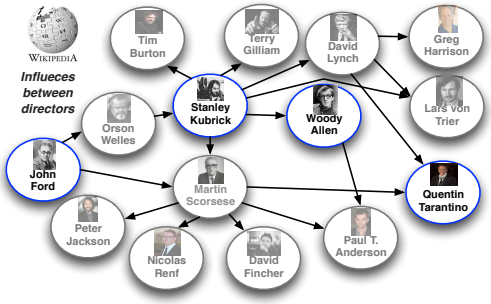

Figure 1: A Web region taken from Wikipedia.

An example. Fig. 1 shows a Web region taken from Wikipedia where the user *Syd* has marked his favorite directors, that is, J. Ford, S. Kubrick, W. Allen and Q. Tarantino. The region besides these nodes also contains other nodes (lighter nodes). A question arises: how should a good map of *Syd's* favorite directors look like? Fig. 2 shows two possible maps. Map 1 contains more nodes and edges than Map 2. Map 2 adopts a specific conciseness strategy: it *minimizes* the number of nodes and edges to keep connectivity among pairs of *distinguished nodes* (i.e., Syd's favorite directors). The node M. Scorsese is not included since it is not a distinguished node, but the connectivity between J. Ford and Q. Tarantino (both distinguished nodes) is still maintained via the direct edge e_2. The edge e_1 in Map 1 is not included in Map 2 because the connectivity between J. Ford and W. Allen is still maintained via S. Kubrick and there is no path in the region going from J. Ford to W. Allen not passing by distinguished nodes.

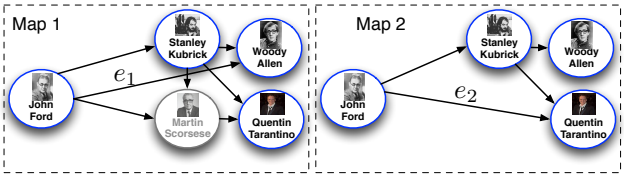

Figure 2: Two possible maps of the region in Fig. 1.

The idea of a map of a region is essentially that of *reflecting* in a *concise way* information in the region in terms of *connectivity* among distinguished nodes. However, how much of the original region has to be included in the map? The writer J. L. Borges scoffs at perfectly accurate maps when he talks about a "map of the Empire...which coincided point for point with it". At the other extreme, *minimal* maps are those that only include nodes with no information about their connectivity (e.g., bookmarks). In between there are maps that besides the distinguished nodes also provide information about their connectivity (e.g., Map 1 and Map 2). A flexible mapping framework should consider different types of maps. In what follows we give the rationale behind our idea of Web maps (Section 2), describe how to map this framework in the current Web infrastructure (Section 3) and then conclude (Section 4).

2 FORMAL MAPS ON THE WEB

The study of making maps is known as cartography [10]. Cartography relies on the human mind's ability to read complex information represented in the map. In the following we provide a formal and general definition of map of a graph where nodes represent objects (e.g., people) and edges relations (e.g., friendship) among them. As we will show, the mathematical characterization of the "object" map brings in both new challenges and opportunities. On one hand, we have to face research questions such as: what is a good map? How to compute efficiently maps? On the other hand, maps can be given a "machine-readable" (e.g., in RDF) representation and then can be shared, exchanged, reused and composed.

2.1 Maps as mathematical objects

We model the Web space as a directed graph. Let $\Gamma = (V_\Gamma, E_\Gamma)$ be a Web Region, where V_Γ and E_Γ are the set of nodes and directed edges, respectively. In the process of map constructions, we ignore edge labels (if present) as we are interested in capturing reachability between nodes. Fig. 1 is an example of Web region. In the remainder of the paper, we use the following notation: $u \to v$ denotes an edge $(u, v) \in E_\Gamma$ and $u \twoheadrightarrow v$ a path from u to v in Γ.

Definition 1 (Map). *A map $M = (V_M, E_M)$ of $\Gamma = (V_\Gamma, E_\Gamma)$ is a graph s.t. $V_M \subseteq V_\Gamma$ and each edge $(x, y) \in E_M$ implies $x \twoheadrightarrow y$ in Γ.*

The idea of a map is essentially that of representing reachability between pairs of *distinguished nodes* (i.e., nodes in V_M), in a concise way. Such definition captures some basic form of map defined over the Web, such as bookmarks. With bookmarks, the set of distinguished nodes is the set of pages in the Web graph that have been marked as interesting. However, this notion of map does not consider reachability information among distinguished nodes.

Definition 2 (Complete Map). *A map $M = (V_M, E_M)$ of $\Gamma = (V_\Gamma, E_\Gamma)$ is complete if, and only if, for all $x, y \in V_M$ it holds that $x \twoheadrightarrow y$ in Γ implies $x \twoheadrightarrow y$ in M, .*

Complete maps address the problem of reachability among distinguished nodes. A possible complete map of the region in Fig. 1 is shown in Map1 in Fig. 2. It includes some direct edges, for instance, between J. Ford and S. Kubrick although not originally present in the region. However, sometimes completeness is not enough to summarize information via maps. The direct edge in the complete map between J. Ford and S. Kubrick is useful because it indicates the fact that S. Kubrick can be reached from J. Ford passing through nodes (O. Welles) not belonging to the map (see Fig. 1). Consider now the edge e_1 in Map 1 in Fig. 2, between J. Ford and W. Allen. Compared to the previous case, this edge does not serve the same purpose. In fact, the connectivity between J. Ford and W. Allen is still maintained via S. Kubrick and there is no other path in the region going from J. Ford to W. Allen only passing for non distinguished nodes. Therefore, e_1 is redundant. Avoiding redundancy is crucial for the purpose of *minimizing* the amount of information necessary to keep connectivity between pairs of distinguished nodes. We need to refine the notion of map. Let $\Gamma = (V_\Gamma, E_\Gamma)$ be a Web region and $N \subseteq V_\Gamma$ a set of nodes. We write $u \twoheadrightarrow_N v$ if, and only if, there is a path from u to v in Γ not passing through any intermediate node in N.

Definition 3 (Good Map). *A map $M = (V_M, E_M)$ of $\Gamma = (V_\Gamma, E_\Gamma)$ is good if, for each pair of nodes $x, y \in V_M$, the following two properties hold:*

(1) $x \twoheadrightarrow_{V_M} y$ in Γ implies $x \to y$ in M;
(2) $x \to y$ in M implies $x \twoheadrightarrow_{V_M} y$ in Γ.

Note that a good map is also a complete map. Map 2 in Fig. 2 shows a good map of the region in Fig. 1. The idea behind the notion of good map is that of giving an economic representation of the reachability between pairs of distinguished nodes. Moreover, good maps have an important property as reported in the following theorem.

Theorem 4. *Let $\Gamma = (V_\Gamma, E_\Gamma)$ be a Web region. Given $N \subseteq V_\Gamma$, there is a unique good map M over Γ with $V_M = N$.*

As discussed in the Introduction, a flexible map framework should consider different types of maps. Accurate maps are the region themselves. Good maps are an example of minimal maps that include connectivity information. However, in some cases one would need to include more nodes besides the distinguished nodes. We now introduce *k-maps*, a family of good maps, which considers nodes in the region having some properties. To this end, let $f : V_\Gamma \to \mathcal{R}$ be a real-valued function defined over the nodes of the region. The function f can be, for instance, a measure of the centrality of nodes (e.g., PageRank) or a popularity measure (e.g., number of incident edges).

Definition 5 (k-maps). *Let $\Gamma = (V_\Gamma, E_\Gamma)$ be a Web region and $f : V_\Gamma \to \mathcal{R}$ be a real-valued function. The k-map of Γ is the good map generated by the set of distinguished nodes $\{v \in V_\Gamma : f(v) \geq k\}$.*

Computing Good Maps. Maps capture information in a region given a set of distinguished nodes. We have the following result.

Theorem 6. *Let $\Gamma = (V_\Gamma, E_\Gamma)$ be a Web region. Given $V_M \subseteq V_\Gamma$, the unique good map $M = (V_M, E_M)$ of Γ can be computed in time:*

(1) $O(|V_M| \times (|V_\Gamma \setminus V_M| + |E_\Gamma|))$ if Γ is a general graph.
(2) $O((|V_M| \times |V_\Gamma \setminus V_M|) + |E_\Gamma|)$ if Γ is a DAG.

2.2 Algebra of Maps

In this section, we research algebraic properties of maps and define operations over them. The following theorem shows the properties of a family of maps.

Theorem 7. *Let $\Gamma = (V_\Gamma, E_\Gamma)$ be a Web region and $\mathcal{M}(\Gamma)$ be the set of all maps over Γ. Furthermore, let $M_i = (V_{M_i}, E_{M_i}) \in \mathcal{M}(\Gamma)$ be maps.*

(1) The binary relation $\sqsubseteq$ over $\mathcal{M}(\Gamma)$, defined by $M_1 \sqsubseteq M_2$ iff $V_{M_1} \subseteq V_{M_2}$, is a partial order on $\mathcal{M}(\Gamma)$.
(2) The order $\sqsubseteq$ induces a Boolean algebra $(\mathcal{M}(\Gamma), \sqcup, \sqcap, \Gamma, \emptyset)$, where: $M_1 \sqcup M_2$ is the unique good map of Γ over $V_{M_1} \cup V_{M_2}$; $M_1 \sqcap M_2$ is the unique good map of Γ over $V_{M_1} \cap V_{M_2}$.
(3) There is an isomorphism of Boolean algebras from $(\mathcal{P}(V), \cup, \cap, V, \emptyset)$ to $(\mathcal{M}(\Gamma), \sqcup, \sqcap, \Gamma, \emptyset)$, given by $N \mapsto M_N$ (the unique good map of N over Γ).

Having well defined operations over maps enables to obtain new maps from other maps. The question is if the re-computation of a map can be (partially) avoided. The next results show this possibility. For a given Web region $\Gamma = (V_\Gamma, E_\Gamma)$ and $S \subseteq V_\Gamma$, we denote by S^*_Γ the transitive closure of S over Γ, i.e., the graph $(S, \{(x, y) : x \twoheadrightarrow_S y \text{ in } \Gamma\})$.

Proposition 8. *Let $M_1 = (V_{M_1}, E_{M_1})$, and $M_2 = (V_{M_2}, E_{M_2})$ be good maps over Γ.*

(1) $M_1 \sqcap M_2 = (V_{M_1} \cap V_{M_2})^*_{M_1} \cup (V_{M_1} \cap V_{M_2})^*_{M_2}$
(2) $E_{M_1 \sqcup M_2} \subseteq E_{M_1} \cup E_{M_2} \cup \{(x, y) \in E_\Gamma : x \in V_{M_1}, y \in V_{M_2}\} \cup$
$\cup \{(y, x) \in E_\Gamma : x \in V_{M_1}, y \in V_{M_2}\} \cup$
$\cup \{x \twoheadrightarrow_{V_{M_1} \cup V_{M_2}} y, x \in V_{M_1}, y \in V_{M_2}\} \cup$
$\cup \{y \twoheadrightarrow_{V_{M_1} \cup V_{M_2}} x, x \in V_{M_1}, y \in V_{M_2}\}$

Corollary 9. *The map $M_1 \sqcap M_2$ can be computed only based on information available in the maps M_1, M_2 and in time $O(|V_{M_1} \cap V_{M_2}| \times (|V_{M_1}| + |E_{M_1}| + |V_{M_2}| + |E_{M_2}|))$. Moreover, the approximation to $M_1 \sqcup M_2$ (modulo redundancy) cannot be computed more efficiently than computing the good map over $V_{M_1} \cup V_{M_2}$ from scratch.*

3 DECLARATIVE SPECIFICATION OF WEB REGIONS

We briefly discuss the problem of how to declaratively specify Web regions and keep information about connectivity among nodes. This need is codified in the following general problem: given a graph $G = (V_G, E_G)$ and a set of nodes $N \subseteq V_G$, construct a subgraph (a region) $R = (V', E')$ of G such that $N \subseteq V'$.

Faloutsos el al. [4] address a variant of this problem: given an edge-weighted undirected graph, two vertices s, t, and an integer k, find a connected subgraph H of size k containing s, t that maximizes a given goodness function. Other approaches have been proposed to discover groups of persons (e.g., [1]) or simplify networks (e.g., [12]). However, these approaches do not provide algebras to manipulate the objects that are produced. Besides, they assume that the whole G is locally available; this hinders their applicability to distributed graphs such as the Web graph.

To formally specify and obtain regions of the Web, we leverage graph navigational languages. A navigational language is a set of functions ("queries") of the form $V_G \to \text{subgraphs}(G) \times \mathcal{P}(V_G)$ that assign to each node v a subgraph (the visited nodes and edges) plus a set of distinguished nodes (the resources selected). Many navigational languages (e.g., XPath, nSPARQL [9]) enable finding pairs of nodes connected by a sequence of edge labels matching some pattern (or navigational expression) expressed via regular expressions over the alphabet of edge labels. This is not enough for our goal; our framework can work with whatever navigational language whose semantics is able to output subgraphs instead of sets of pairs of nodes (such as NautiLOD [5, 6, 8]).

3.1 The Implemented System

The framework to build maps of the Web of Linked Data has been implemented in Java in the MaGe tool[1] and uses NautiLOD to specify and construct Web regions. We discuss now a real-world example[2].

[1] which can be downloaded at the address http://mapsforweb.wordpress.com
[2] Data from 2016.

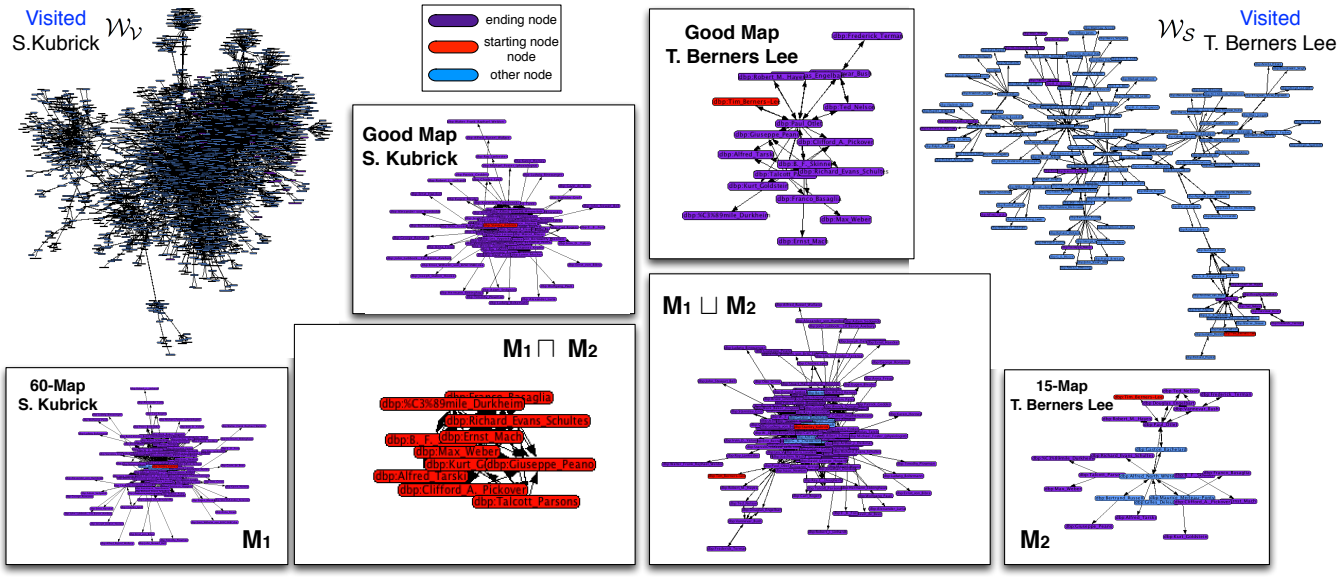

Figure 3: Influence maps of S. Kubrick and T. Berners Lee only considering scientists up to distance 6.

Example 10. ***(Maps of influence networks and algebra)*** *Specify two regions that contain people that have influenced or have been influenced up to distance 6 by Stanley Kubrick (SK) or Tim Berners-Lee (TBL). The ending nodes in the regions must be scientists. Compute maps and use the algebra of maps.*

The region associated to the influence network of SK contains 2981 nodes and 7893 edges. The good map associated to SK (109 nodes; 2629 edges) summarizes the region and then provides insight on the connectivity between ending nodes (i.e., scientists that have been influenced or have influenced SK) and with SK. We zoomed in this influence path by computing the 60-map (M_1) of the region (120 nodes; 3627 edges).

The region associated to TBL is smaller (149 nodes; 236 edges). The associated good map (18 nodes; 43 edges) tells us, for instance, that there exists an influence path from TBL to G. Peano passing via P. Outlet. When zooming in this path, by computing the 15-map (M_2) of the region (23 nodes; 43 edges), we discovered that the non ending node B. Russell is also in the path.

Fig. 3 also shows examples of the algebra of maps. It shows the intersection between M_1 and M_2. The result is the good map that could have been obtained by making the union of the regions and then computing the good map from the set of distinguished nodes (see Definition 3) given by $V_{M1} \cap V_{M2}$.

However, the advantage of using the algebra is to avoid to compute from scratch the good map and obtain it without looking at the regions. As an example, in the intersection of M_1 and M_2 we have the nodes G. Peano and A. Tarski, which means that both belong to the influence networks of SK and TBL. The map of the union of M_1 and M_2 enables to put together information from the two maps.

This enables to discover possible additional influence relations between pairs of nodes that are not present in the two maps. In this specific example, there is no path between SK and TBL neither in M_1 nor in M_2. However, the union of the k-maps enabled to discover the connection between TBL and SK (i.e., TBL→P. Outlet→B. F. Skinner←SK).

4 CONCLUDING REMARKS

Due to limitations of human I/O capabilities, the management of information at a Web scale calls for automatic mechanisms and thus machine-processable information. In this paper we have shown that maps, key devices in helping human navigation in information spaces, are meaningful on the Web space. We think that the formal models presented here are a starting point for further developing of cartography on the Web.

REFERENCES

[1] J. Adibi, H. Chalupsky, E. Melz, A. Valente, et al. The KOJAK Group Finder: Connecting the Dots via Integrated Knowledge-based and Statistical Reasoning. In *AAAI*, pages 800–807, 2004.
[2] M. Dodge and R. Kitchin. *Atlas of Cyberspace.* Addison-Wesley Great Britian, 2001.
[3] P. Doemel. WebMap: a Graphical Hypertext Navigation Tool. *Computer Networks and ISDN Systems*, 28(1):85–97, 1995.
[4] C. Faloutsos, K.S. McCurley, and A. Tomkins. Fast Discovery of Connection Subgraphs. In *KDD*, pages 118–127. ACM, 2004.
[5] V. Fionda, C. Gutierrez, and G. Pirró. Extracting Relevant Subgraphs from Graph Navigation. In *ISWC (Posters & Demos)*, volume 914. CEUR-WS.org, 2012.
[6] V. Fionda, C. Gutierrez, and G. Pirró. Semantic Navigation on the Web of Data: Specification of Routes, Web Fragments and Actions. In *WWW*, pages 281–290. ACM, 2012.
[7] V. Fionda, C. Gutierrez, and G. Pirrò. Building knowledge maps of web graphs. *Artif. Intell.*, 239:143–167, 2016.
[8] V. Fionda, G. Pirrò, and C. Gutierrez. Nautilod: A formal language for the web of data graph. *TWEB*, 9(1):5:1–5:43, 2015.
[9] J. Pérez, M. Arenas, and C. Gutierrez. nSPARQL: A Navigational Language for RDF. *JWS*, 8(4), 2010.
[10] A. H. Robinson, J. Morrison, O. C. Muehrcke, A.J. Kimerling, and S. C. Guptill. *Elements of Cartography.* Wiley, 1995.
[11] D. Shahaf, C. Guestrin, and E. Horvitz. Trains of Thought: Generating Information Maps. In *WWW*, pages 899–908. ACM, 2012.
[12] F. Zhou, S. Malher, and H. Toivonen. Network Simplification with Minimal Loss of Connectivity. In *ICDM*, pages 659–668. IEEE, 2010.