Enriching and Cutting: How to Visualize Networks Thanks to Linked Open Data Platforms

Léa Saint-Raymond, Antoine Courtin

Artl@s Bulletin

Volume 6
Issue 3 *Visualizing Networks: Approaches to Network Analysis in Art History*
Article 7

2017

Enriching and Cutting: How to Visualize Networks Thanks to Linked Open Data Platforms.

Lea Saint-Raymond
ENS / Université Paris Ouest Nanterre La Défense, lea.saint-raymond@ens.fr

Antoine Courtin
Institut national d'histoire de l'art, antoine.courtin@inha.fr

Enriching and Cutting: How to Visualize Networks Thanks to Linked Open Data Platforms.

Cover Page Footnote

A first version of this paper was presented in 2015 at the Faculdade de Ciencias Sociais e Humanas, in Lisbon, during the International Conference of the Historical Network Research. We thank all the participants for their enlightening comments on our work.

Enriching and Cutting: How to Visualize Networks Thanks to Linked Open Data Platforms.

Léa Saint-Raymond *
Université Paris Nanterre / ENS
Antoine Courtin **
INHA

Abstract

Conspicuous in the social sciences, networks analyses are becoming more common in art history. This paper takes a pragmatic look at network visualizations through the study of a specific corpus: Parisian auction sales for "modern paintings." This initial data set was enriched using linked open data platforms and technologies for realigning datasets, and then visualized to represent networks of artists and buyers. Although these visualizations bring about an overview of the market and allows very close readings, they run the risk of being illegible, unless they are combined with other modes of visualization.

Résumé

Après s'être développée en sciences sociales, l'analyse de réseaux prend son essor en histoire de l'art. Cet article adopte un point de vue pragmatique sur cette méthode de visualisation, à partir d'un jeu inédit de données : les ventes aux enchères publiques parisiennes de « tableaux modernes ». Ce corpus initial a été enrichi grâce au web sémantique et aux plates-formes de données ouvertes et liées, puis représenté par des réseaux d'artistes et d'acquéreurs. Bien que permettant de dégager une vue d'ensemble du marché, ce mode de visualisation risque néanmoins de s'avérer peu lisible, s'il n'est pas combiné à d'autres outils plus traditionnels de représentation des données.

** Léa Saint-Raymond is a Ph. D. candidate in art history, at the Université Paris Nanterre. Alumna of the École normale supérieure, Paris, she is agrégée in economic and social sciences.*
*** Antoine Courtin supervises the documentary engineering at the Institut national d'histoire de l'art. He received a double training in art history and new technologies applied to historical sciences.*

Introduction

Quantitative work related to network analyses in art history has been developing for over a decade thanks to Béatrice Joyeux-Prunel's pioneer works, the symposium she organized in 2008[1] and the Artl@s project. Within the discipline of art history, network visualization is growing, supported by new dissertations,[2] by Pamela Fletcher's and Anne Helmreich's research[3] and by innovation-friendly journals, such as *Nineteenth-Century Art Worldwide* and the *Artl@s Bulletin*. In particular, networks called into question the supposed centrality of Paris as the capital city of Modernism in the interwar period.[4] They have also helped in analyzing the artistic field of the portraitists in the late 19th century and in the interwar period.[5] Networks were used in the social sciences in order to understand the artistic field in Europe from the late 19th century[6] through the 1930s.[7] Nevertheless, the exciting prospect of using a new tool can turn out to be unsatisfactory if some precautionary principles are bypassed. This paper will adopt a pragmatic and methodological approach in order to see the promises and problems of network visualization.

Based on a case study, we will get our hands dirty in order to build some meaningful networks and to see their limits. First, we will describe our initial corpus regarding the auction market for "modern paintings" in Paris and taking part in Léa Saint-Raymond's research.[8] The second step is thus to enrich this initial corpus, thanks to the available linked open data platforms.[9] As a matter of fact, "networking the networks" has become possible with the semantic web.[10] By giving some extra information about the nodes of the dataset, this enrichment is the only way to avoid the horizontal character of networks and, thus, to create a fuller picture of the social fields. The final step of the process is to build these enriched networks. Visualizing them is a process that has evolved from multiple disciplines, the geometry of statistics, spatialization of data, surfaces and signifying relationships, etc. Building networks has apparently been streamlined by the new visualization software. However, especially in the social sciences and humanities, the process often remains artisanal, with many iterative cycles and with many opportunities for adjustment. Aware of the force of graphical forms on the production of ideas,[11] we have to discuss the reality traced by networks, asking whether they are the best way to analyze the history of artistic exchanges. Too often, the best is the enemy of the good: too much information and too many nodes may represent a handicap, hence the need for both enriching and cutting the networks.

The initial dataset and its need for contextualization

One issue of Léa Saint-Raymond's research is to understand the structure of the art market regarding living artists. To do so, networks constitute a powerful tool, and the data, the main problem. As a matter of fact, analyzing the artistic networks for the "primary art market" is a real challenge due to the impossibility of having a comprehensive accessibility to the account books for the art dealers, the collectors and the artists. The only possible networks can be drawn for all the

[1] Béatrice Joyeux-Prunel (éd), *L'Art et la Mesure. Histoire de l'art et approches quantitatives : sources, outils, méthodes*, proceedings of the symposium held at the Ecole normale supérieure, December 3-5, 2008, (Paris: Editions de la rue d'Ulm, 2010).
[2] Matthew Lincoln, "Modeling the Dutch and Flemish Print Production Network, 1550-1750", University of Maryland, College Park, defended in 2016.
[3] Pamela Fletcher and Anne Helmreich, "Local/Global: Mapping Nineteenth-Century London's Art Market", *Nineteenth-Century Art Worldwide*, Volume 11, Issue 3 (Autumn 2012).
[4] Béatrice Joyeux-Prunel, "Provincializing Paris. The Center-Periphery Narrative of Modern Art in Light of Quantitative and Transnational Approaches", *Artl@s Bulletin* 4, no. 1 (2015): Article 4.
[5] Léa Saint-Raymond, "Bas les masques ! Pour une relecture socio-économique du Montparnasse des années 1920", *Artl@s Bulletin* 4, no. 2 (2016): Article 5.
[6] Robert Jensen, *Marketing Modernism in fin-de-Siècle Europe* (Princeton: Princeton University Press, 1994).
[7] Fabien Accominotti, *Le marché de la peinture moderne à Paris, 1900-1930 : une affaire en termes de sociologie économique,* Ph. D dissertation (Paris, École des Hautes Études en Sciences Sociales, 2010).
[8] Léa Saint-Raymond, *Le pari des enchères : le lancement de nouveaux marchés artistiques à Paris entre les années 1830 et 1939*, Ph. D. dissertation, supervised by Ségolène Le Men, Université Paris Nanterre.
[9] Linked open data platforms can be defined as a collaborative movement led by W3C, which promotes common methodologies for data exchange. This movement is based on the web of data, which links and structures information on the Internet. It thus relies on several technological bricks such as the URIs, the RDF model, OWL ontology and the SPARQL request protocol.
[10] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", Scientific American, May 2001, p. 29-37.
[11] Johanna Drucker, *Graphesis: Visual Forms of Knowledge Production* (Harvard: Harvard University Press, 2014).

"common artists" between art galleries,[12] but they miss crucial information: prices. However, artistic networks and prices are easier to observe through the prism of auction sales, which belong to the "secondary art market". This was the starting point for the constitution of Léa Saint-Raymond's initial dataset in 2014.

This corpus is very much inspired by the Sales Catalogs database of the Getty Provenance Index,[13] currently headed by Christian Huemer at the Getty Research Institute. According to the September 2015 figures, this online dataset presents 1,129,381 records of British, Belgian, Dutch, Scandinavian, French and German auction catalogs, spanning mainly from 1680 through 1840 for the first five categories, and from 1930 through 1945 for the German sales.[14] The Sales Catalogs database lists all the information about the auction sale — auction house, city and country of sale, names of the auctioneers and the experts — as well as the artworks that were sold, including the name of the artist, title and description of the lot. When annotated, the catalog also records the buyer and seller names, and the hammer prices. Likewise, our own dataset is built on the auction catalogs, but focuses on the Parisian sales for "modern painting" ("*tableaux modernes*") from the 1850s through the 1930s. Contrary to the Getty Provenance Index, the minutes of the auction sales ("*procès-verbaux*") were used to get access to the hammer prices and the identity of both purchasers and sellers. Curated at the Archives de Paris, these administrative documents present the advantage of being unambiguous, whereas two annotated catalogs of the same auction sale may be contradictory because they rely on the eye and the ear of the beholder.

From this initial dataset, the auction market visualization could already be obtained, such as the one provided by Maximilian Schich for the Getty Provenance Index, showing the network diagram of agents connecting the British, Belgian, Dutch and French auction markets from 1801 through 1820.[15] With this example, one could build a network that would display the purchasing and selling operations of the agents based on the address given in the minutes. Nevertheless, our corpus is still too flat to produce a similar result. Indeed, the majority of catalogs gave very little information about the artist, unlike the contemporaneous ones. In the 19th century, very few Parisian catalogs provided their date and place of birth or death, or their nationality. In addition, the artist's first name was not systematically recorded. Since our dataset is based on the primary sources, it considers that "Bonheur (R.)", "Bonheur" and "Bonheur (Rosa)" stand for three different artists, whereas they are one and the same person, Marie-Rosalie Bonheur, called Rosa Bonheur. Before any network visualization, we thus have to clean up the data, *i.e.* to add an "authority record" to the "verbatim" name of the artist. By doing so, we will thus enrich our information.

How to enrich data thanks to linked open data platforms.

Enriching a dataset can be done manually and, most of the time, it is the only way to do so. For instance, when the minute of the sale records the Parisian address of the purchaser, the buyer's identification passes through the *Bottin du commerce*. This annual address book registers Parisian residents in three different ways: in alphabetical order, in occupation order, or by the list of streets. The *Bottin du commerce* constituted a comprehensive source to study the Parisian art dealers and their geography from 1815 through 1955,[16] and it is also very useful to give authority names for most of the actors in our dataset. Regarding the painters and sculptors whose artworks were sold at auction, a comprehensive resource helps identify them: the

[12] Fabien Accominotti, *Market Chains: Careers and Creativity in the Market for Modern Art* (Princeton, Princeton University Press, to be published).
[13] http://www.getty.edu/research/tools/provenance/search.html, accessed January 5, 2017.
[14] http://www.getty.edu/research/tools/provenance/charts.html, accessed January 5, 2017.
[15] http://www.getty.edu/research/tools/provenance/zoomify/index.html accessed January 5, 2017.
[16] Léa Saint-Raymond, Félicie de Maupeou and Julien Cavero, "Les rues des tableaux: The Geography of the Parisian Art Market. 1815-1955", *Artl@s Bulletin* 5, no. 1 (2016): Article 10.

Benezit Dictionary of Artists.[17] Its publication dates from 1911 and it provides the largest resources for artists' biographies. Unfortunately, the online version of the Benezit dictionary doesn't provide open data.[18] As a consequence, the enrichment cannot be done automatically with dynamic and automatic queries, hence the need for manual entries. For the researcher, this operation can better situate the context and the artists whose artworks were sold at auction. However, manual data enrichment may seem quite discouraging for the analyst when dealing with a huge corpus, unless we find a digital resource in linked open access.

Fortunately, the research environment has been changing dramatically. In recent years, we witnessed many web-based initiatives for cultural institutions, in particular the advent of open data and the maturity of semantic web technologies. In addition, digitized contents are becoming more and more available on the web. They require associated metadata in order to gain visibility. In the end, a very active community of researchers and institutions aims at connecting data in human and social sciences, *i.e.* linking datasets to other datasets. All these initiatives can be brought together under two acronyms: OpenGLAM[19] and LODLAM,[20] the latter standing for "Linked Open Data in Libraries, Archives and Museums". The challenge of LODLAM may be illustrated with a metaphor. Beforehand and still today, researchers used to store data in their own "grain silos", which were protected and quite closed. LODLAM endeavors to link datasets allowing automatic processing, or in other words, to interconnect these silos. In parallel with this data opening, LODLAM promotes licenses to reuse this data. At the moment, the community has been able to publish a huge number of datasets in linked data format.[21]

The aim was thus to enrich automatically Léa Saint-Raymond's initial dataset by using a linked open data platform (Fig. 1). For the artists' biographies, the Union List of Artist Names (ULAN) meets the criteria.[22] Provided by the Getty Research Institute, this Linked Open Data thesaurus contains around 293,000 different names, pseudonyms and variant spellings corresponding to 120,000 artists, but gives also their nationality, their gender, their roles, birth and death places, and events related to them. For instance, thanks to this thesaurus, we learn that R. Bonheur's authority name in ULAN is "Bonheur, Rosa", that she was a French woman and an animal painter, that she was born in Bordeaux in 1822 and died in Thomery in 1899. We also used Wikidata[23] to enrich the artist's identity. This source is, of course, very questionable for an audience of historians, but, compared to ULAN, more artists were found in this linked database.

Once we have found these open data platforms, we can link them to the initial corpus in order to enrich it. Although they are closed lists of historical actors and not free texts, the technologies we use remain related to Named Entity Recognition (NER). The NER is a subtask of information extraction; it seeks to locate and classify named entities from a text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Disambiguation is the focal point of entities recognition and, above all, the issue of any realignment. A research field called NLP (Natural language processing) deals with this challenge and, more generally, studies the interaction between computers and human languages, as underlined in Grishman and Sundheim's work.[24]

The purpose of extracting and then realigning entities of historical actors (named entities grouping other typologies such as organizations, time concepts up to product names, events, etc.) is

[17] Emmanuel Bénézit, *Dictionnaire critique et documentaire des peintres, sculpteurs, dessinateurs et graveurs de tous les temps et de tous les pays…*, 14 vol. (Paris: Gründ, 2006).
[18] http://www.oxfordartonline.com/public/book/oao_benz, accessed January 5, 2017.
[19] http://openglam.org/principles/, accessed January 5, 2017.
[20] http://lodlam.net/, accessed January 5, 2017.
[21] http://lod-cloud.net/, accessed January 5, 2017.
[22] http://www.getty.edu/research/tools/vocabularies/ulan/, accessed January 5, 2017.
[23] Wikidata is a project of the Wikimedia Foundation: a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other Wikimedia projects, https://www.wikidata.org/wiki/Wikidata:Introduction.
[24] Ralph Grishman and Beth Sundheim, "Message Understanding Conference: A Brief History" in *Proceedings of the 16th Conference on Computational Linguistics, Vol. 1,* (Stroudsburg: Association for Computational linguistics, 1966): 466-471.

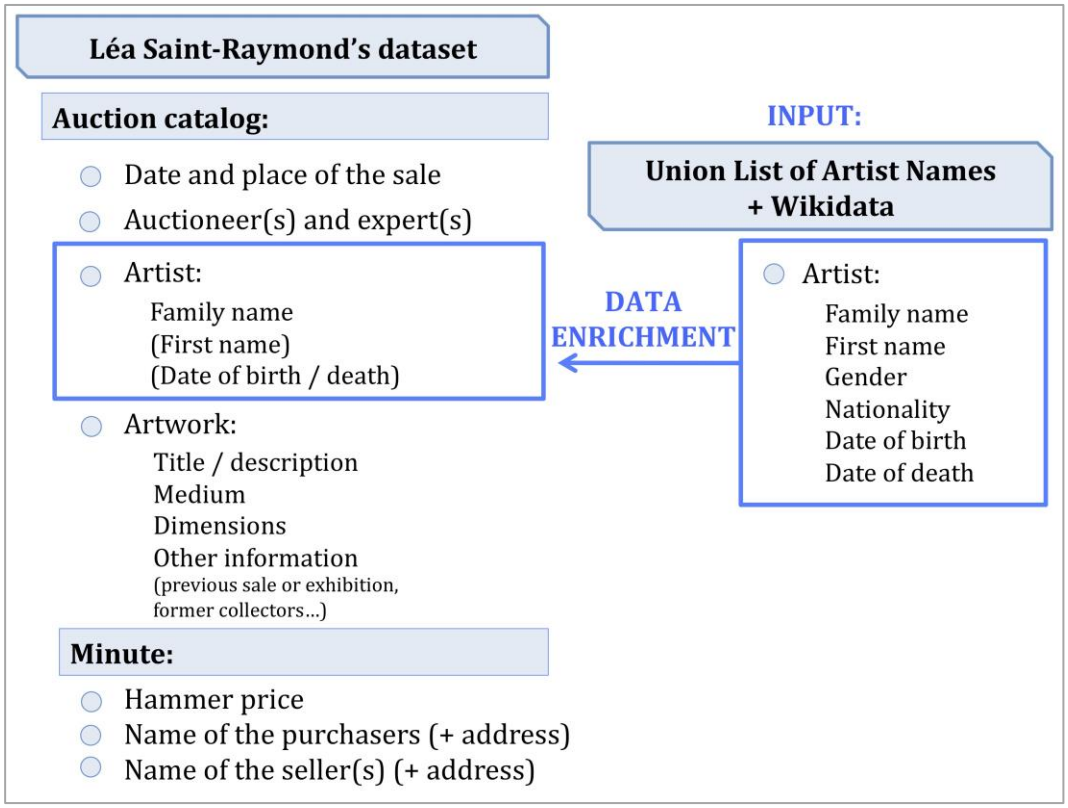

Figure 1. Protocol of data enrichment, from our initial dataset.

threefold: to analyze the input data in order to detect entities, to assign them a type (in our case, historical actors), and to propose a list of Uniform Resource Identifiers[25] (URIs) from third-party services. The assignment is based on a so-called *confidence score*.

A decision interface of realignment is then used to interconnect all the open vocabularies through the web, such as ULAN and Wikidata.[26] This interface comes from tools named *Data Transformation Tools* or *Self-service data preparation*, whose purpose is to manipulate, clean and interconnect datasets of various natures and sources. These tools have the advantage of being similar to an Excel spreadsheet that would have been boosted. This facilitates the appropriation of the tool and its methodology by people closest to "business" data and not only by data scientists or information computing specialists. Among these Integrated Device Technologies (IDTs), the most widely used ones are currently Wrangler, from Stanford University, and OpenRefine (initially Freebase Gridworks and then Google Refine), but paying solutions exist such as Trifacta, Talend data preparation, Dataiku DSS, *etc*. In the field of Digital Humanities, OpenRefine has rapidly become popular, and we chose it for this experimentation. Although the sustainability of its development is often questioned, it reconciles data with existing knowledge bases and enables the principles of linked data thanks to increased behavior and plugins.[27]

[25] Tim Berners-Lee (July 2002). "What Do HTTP URIs Identify?". Internet Engineering Task Force. Retrieved 15 January 2007. The URI was originally called a UDI, and originally all URIs identified information objects. Now, URI schemes exist that identify more or less anything.

[26] S. Van Hooland, R. Verborg, M. de Wilde, J. Hercher, E. Mannens, R. Van de Walle, "Evaluating the success of vocabulary reconciliation for cultural heritage collection", in *Journal of the American Society for information Science and Technology*, 2013, vol. 64, n°63: 464-479.

[27] See in particular http://refine.codefork.com , http://refine.deri.ie/ and http://freeyourmetadata.org/named-entity-extraction/ , accessed January 5, 2017.

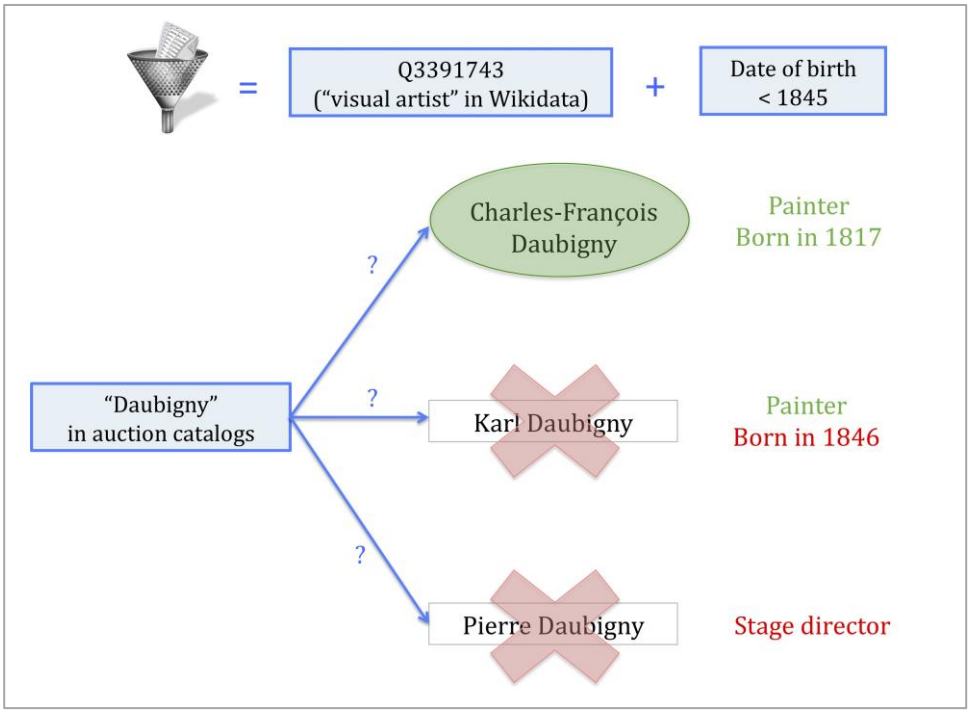

Figure 2. The reconciliation of "Daubigny".

More specifically, we used OpenRefine's RDF Refine Extension. RDF stands for Resource Description Framework and it is a powerful tool of the semantic web. As a matter of fact, through RDF extension, it is possible to make dynamic and automatic queries. We decided to test this protocol for a part of the corpus regarding all the auction sales for "modern paintings" that took place in Paris in 1868. Compared to the whole dataset, which records 44,630 artworks in January 2017, it is a "small" sample, containing 20 auction sales, 286 artists and 1,100 artworks. This choice was motivated by the study of the painter Félix Ziem's account book: as he decided to sell his watercolors at auction in Paris in 1868, there was a need to compare his position with that of other artists in the Parisian auction sales of the period. The outcome of this research — and the networks of the Parisian auction market in 1868 — was published in 2016.[28] This paper aims to investigate the methodology used to build these networks, rather than analyze the art market in 1868, which can be found in the 2016 article.

In order to enrich Léa Saint-Raymond's initial dataset, we first tried an exact match, restricting the reconciliation to the name of the artist as it was written in the auction catalogs. For instance, "Delacroix (Eugène)", in the 1868 catalog, was reconciled to "Delacroix, Eugène" in ULAN. The outcome was quite disappointing because this exact match was made for only 35 percent of the artists. We thus realized the need for disambiguation. Although the enrichment can be done very quickly with OpenRefine, some precautions must be taken in order to avoid identification mistakes. We had to use several passes to improve the reconciliation. After considering exact matches, we then considered approximate ones (Fig. 2).

For instance, several auction catalogs mention a painter called Daubigny, but the reconciliation process gives three persons whose last name is Daubigny: Charles-François, Karl and Pierre. How can we solve this ambiguity? We decided to code a first pass, removing the persons whose occupation

[28] Léa Saint-Raymond, "How to Get Rich as an Artist: The Case of Félix Ziem – Evidence from His Account Book from 1850 through 1883", *Nineteenth-Century Art Worldwide*, Vol. 15, No. 1 (Spring 2016).

were different from "visual artist": that's why we eliminated Pierre Daubigny, who was a stage director. But we still had to decide whether Daubigny was Charles-François or Karl or to suspend our judgment. We then ran a second pass, considering the date of the auction. For instance, if the sale happened in 1865, it was very likely to include artists whose age would have been above twenty, thus born before 1845. According to this criterion, "Daubigny" wouldn't be the young Karl Daubigny, but Charles-François, who was born in 1817.

Finally, we used a last filter based on proximity by, for example, Key Collision Methods. For instance, if Daubigny is written with a "i" and not a "y" in the catalog, the match is still Charles-François Daubigny. The aggregation of these three passes led to a powerful disambiguation: 73 percent of the artists were identified. It means that one artist over four remained unknown. It is essential to point out that the realignment rate improves over time, thanks to collaborative contributions, particularly in Wikidata. Thus, by replaying the process in January 2017, new realignments have been proposed.

After this operation, we finally exported all the information related to the identified artists: nationality, gender, date and places of birth and death. In other words, we added several columns to our initial .xls file after CSV format transformation.

Visualizing... and cutting

The term "network" refers to three different objects: a social network, a mathematical graph and finally the relational data, that is to say the digital files that list both the elements and the links between them, and that are described by metadata.[29] Three categories of networks can be visualized from our enriched corpus, depending on the persons defining the "nodes": the artists (Figs. 3-5), the purchasers and sellers (Fig. 6), or all these categories at the same time (Fig. 7). Let us consider the networks displaying the artists (Figs. 3-5). The extra information we added, thanks to ULAN and Wikidata, takes the form of a node file with extra-attributes. In our visualization with Gephi, the enrichment is represented through (Figs. 3-5), with the color of the disk. In the first network (Fig. 3), the gender of the artist is represented in blue or red, but the node can also be colored according to the artist's "most commonly associated nationality"[30] (Fig. 4 and 5), or according to the fact of being dead or alive at the moment of the sale. Another layer of information can be added to the node through the size of the node. Here, we decided to map the area of the disk proportionally to "average hammer price". In 1868, the highest one — 25,000 francs — was reached by Henriette Browne, called Henriette Brown at that time: the node that corresponds to her in the network is the biggest (Figs. 3-5). These two layers of information — size and color of the node — embed the network in a broader socio-economic environment. The edges between nodes may represent two different relationships. In the first two networks, two artists are linked together if their artworks were sold in the same auction sale. The more sales in common, the darker the edge (Figs. 3 and 4). Artists can also be connected through their buyers: in the third network (Fig. 5), the edges correspond to the "common purchase" of artworks by two different buyers. In this visualization, the nodes are more scattered when the purchaser focuses his or her attention on one single artist. This relationship has been chosen in Fig. 6, which maps the purchasers as the nodes, and the common purchase by different buyers of artworks from a same artist, as the edge: the more artists they shared, the darker the link. The last network (Fig. 7) reveals all these connections, also displaying both the purchasers and the artists "in common" as the nodes.

[29] Tommaso Venturini, Mathieu Jacomy and Débora Pereira, *Visual Network Analysis* (working paper, 2014) http://www.tommasoventurini.it/wp/wp-content/uploads/2014/08/Venturini-Jacomy_Visual-Network-Analysis_WorkingPaper.pdf, accessed January 5, 2017.

[30] http://www.getty.edu/research/tools/vocabularies/ulan/about.html, accessed January 7, 2017.

Of course, the spatialization algorithm matters a lot: the choice of Force Atlas 2, rather than Fruchterman Reingold[31], for instance, produces similar networks in content but different mappings from a visual point of view. However, with Gephi, the use of the "Noverlap" function, which prevents the edges to overlap, produces little difference between a Force Atlas 2 network (Fig. 5) and a Fruchterman Reingold one (Fig. 8). Beyond these variants, the principles remain the same to know that the nodes repel each other while the links attract them. At first sight, what can be read in our networks are the structural voids and the aggregates, which can be interpreted as communities or as sets sharing common characteristics. For these aggregates, metadata will come into play: is there a correlation between the metadata — the color of the node — and the mapping — the structure of the network? When the metadata are scores rather than categories, the strategy is the same and one can use the node sizes as visual variables. Thanks to a typology measure available by default in Gephi called "betweenness centrality", one can locate the nodes playing particular roles. As a consequence, the hypotheses that are generated relate to the identity of the nodes that play a key role in the structure and the categories, or metrics, that explain them.

The advantage of network visualization is its comprehensiveness. Like a microcosm, one single diagram displays a representation of the auction market in 1868 and allows both distant and close readings. On the one hand, network visualization gives a very direct and efficient overview of the trends. In this particular example, we see that the auction market in 1868 was dominated by a small number of painters, whose high valued artworks constituted the core of the collections (Fig. 7): Théodore Rousseau, Alexandre Gabriel Decamps, Jules Dupré, Eugène Delacroix, Félix Ziem, Narcisse Virgile Diaz de la Peña, Camille Corot, Constant Troyon and Eugène Fromentin. On the other hand, the use of networks can allow a very precise interpretation of the data by picking an artist in particular and looking at his or her position in the market. For instance, Auguste Toulmouche's and Louis Gallait's paintings were both expensive (Fig. 4) and peripheral in the purchases (Fig. 5), whereas Théodore Gudin's artworks entered in several collections at a very low average hammer price. Since all the nodes are put on the same level, network visualization has the great power of "recanonizing" the artists who were successful but who have been forgotten in art history, like Hendrick Leys and Edmond Tschaggeny, two Belgian painters who were the key figures of the Parisian market in 1868 (Fig. 4). As a consequence, network visualization is a powerful tool to make new questions emerge: what was the "taste" of the period? why were some famous artists neglected by history? The famous art historian Francis Haskell raised these issues[32] thanks to network visualization; it is easier to deal with large datasets and to provide some preliminary answers. In addition, without the network of the buyers (Fig. 6), it would have been very difficult to perceive the centrality of three art dealers – Brame, Petit and Binant – on the market for "modern paintings". Networks thus constitute a first step to formulate research questions that will be tested separately by other means, in particular by combining alternative measures and qualitative work.

Nevertheless, the advantage of network visualization is also its main drawback: as the dataset gets bigger, its interpretation is harder. With 286 nodes, our artistic networks can rightly be accused of illegibility. Cutting the sample may thus be an answer in order to better visualize it. The first two networks (Figs. 3 and 4) can be simplified by removing all the "singletons", *i.e.* the artists who just appeared in one auction sale.[33] Even more directly, it is possible to enhance the core of the network thanks to the adjacency matrix.

[31] These Force-directed graph drawing algorithms are a class of algorithms for drawing graphs in an aesthetically pleasing way.

[32] Francis Haskell, *Rediscoveries in Art. Some Aspects of Taste, Fashion and Collecting in England and France* (London: Phaidon Press Limited, 1976).

[33] See Table 10 in Léa Saint-Raymond, "How to Get Rich as an Artist: The Case of Félix Ziem – Evidence from His Account Book from 1850 through 1883", *Nineteenth-Century Art Worldwide*, Vol. 15, No. 1 (Spring 2016).

Artist #1	AND	Artist #2	Number auction sales in common
Diaz de La Peña (Narcisse)	and	Ziem (Félix)	16
Diaz de La Peña (Narcisse)	and	Rousseau (Théodore)	15
Rousseau (Théodore)	and	Ziem (Félix)	14
Diaz de La Peña (Narcisse)	and	Dupré (Jules)	13
Dupré (Jules)	and	Rousseau (Théodore)	13
Decamps (Alexandre Gabriel)	and	Rousseau (Théodore)	12
Dupré (Jules)	and	Ziem (Félix)	12
Corot (Camille)	and	Diaz de La Peña (Narcisse)	11
Decamps (Alexandre Gabriel)	and	Diaz de La Peña (Narcisse)	11
Delacroix (Eugène)	and	Diaz de La Peña (Narcisse)	11
Delacroix (Eugène)	and	Dupré (Jules)	11
Delacroix (Eugène)	and	Rousseau (Théodore)	11
Diaz de La Peña (Narcisse)	and	Troyon (Constant)	11
Dupré (Jules)	and	Fromentin (Eugène)	11
Troyon (Constant)	and	Ziem (Félix)	11
Corot (Camille)	and	Delacroix (Eugène)	10
Corot (Camille)	and	Rousseau (Théodore)	10
Decamps (Alexandre Gabriel)	and	Dupré (Jules)	10
Decamps (Alexandre Gabriel)	and	Ziem (Félix)	10
Delacroix (Eugène)	and	Fromentin (Eugène)	10
Delacroix (Eugène)	and	Ziem (Félix)	10
Diaz de La Peña (Narcisse)	and	Fromentin (Eugène)	10
Fromentin (Eugène)	and	Rousseau (Théodore)	10
Rousseau (Théodore)	and	Troyon (Constant)	10

Table 1. Detail of the adjacency matrix regarding the networks of Figures 2 and 3.

This .xls file counts the number of connections between the nodes. By picking the highest values of edges (Table 1), one can have direct access to the most central artists in the 1868 auction market.

Although network visualization constitutes a new and exciting methodological approach, some older tools seem even more powerful to picture very specific aspects of the market. Instead of the networks in Fig. 3, why not say that, statistically, less than 2 percent of the artists whose artworks were sold at auction were women? A map showing that 52 percent of the artists were French, 6 percent came from the Netherlands and 5 percent from Belgium, etc., could also replace the network of Figure 4.

Finally, the most valued artists in 1868 can be highlighted with the help of a very simple graph (Fig. 9). In order to get the most comprehensive and precise picture of the art market, the use of networks must be combined with other modes of visualization. As Jeremy Boy or Katy Börner points out, it is necessary to think about "literacy in data visualization".[34]

Networks must be handled with care as they do not constitute a purely illustrative picture, but a set of graphic codes that the reader must be able to interpret. In addition, above a threshold limit of nodes, the picture tends to be unreadable. When it comes to dynamic networks with a huge amount of

[34] Jérémy Boy, *Engager les Citoyens à Aller au-delà des Simples Représentations de Données Ouvertes* [unpublished dissertation], Telecom ParisTech, defended in 2015. http://jyby.eu/phd_dissertation/2015_ENST_0025_Jeremy_Boy.pdf , accessed January 5, 2017. The American Library Association defines "information literacy" as a set of abilities requiring individuals to "recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information".

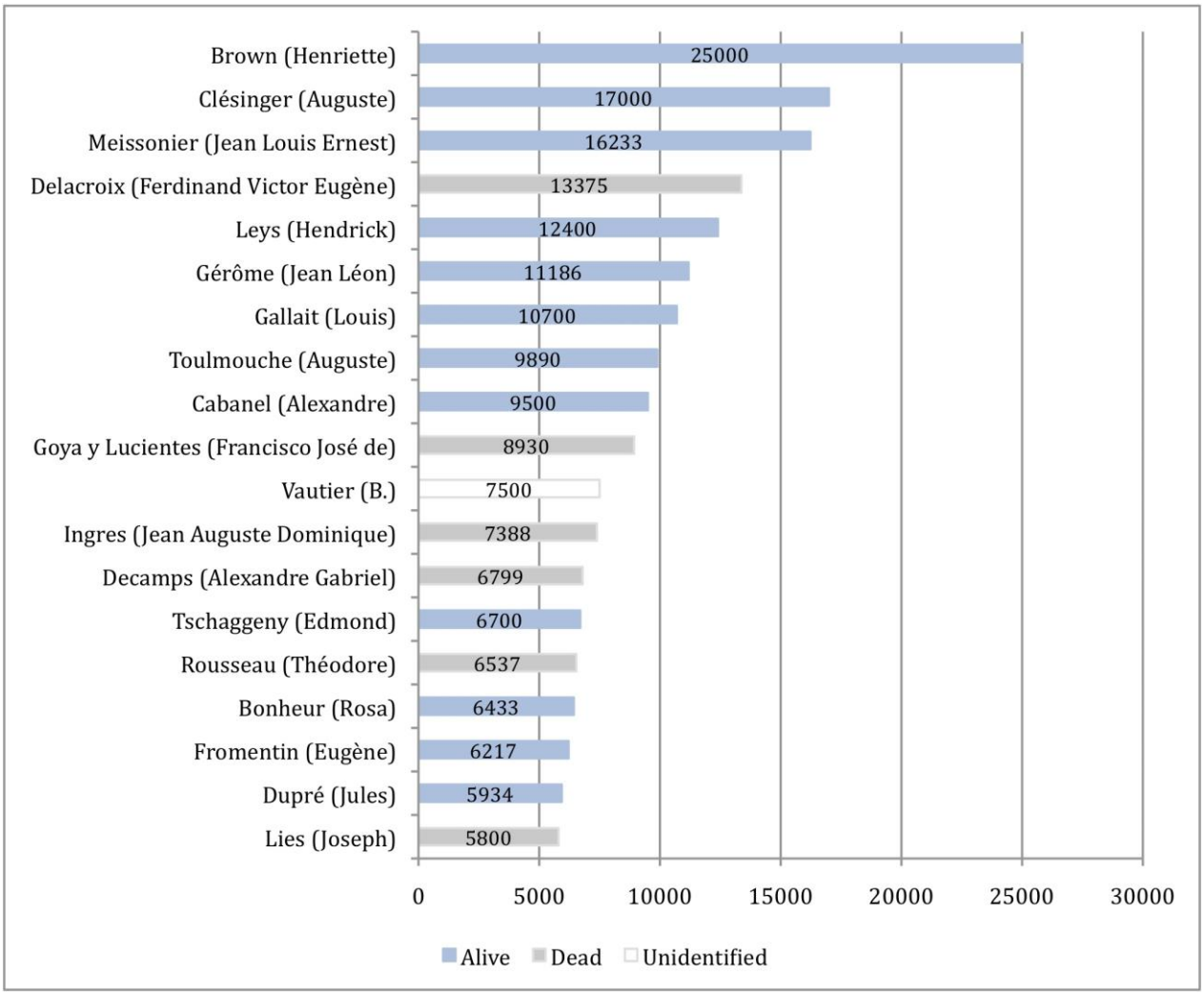

Figure 9. List of the most valued artists in 1868 sorted by the average hammer price.

data, the interpretation is even harder, as exemplified by The Getty Provenance Index network, regarding 230,000 records of agents, connecting the European auction market from 1801 through 1820.[35] In order to take this time variable into account, it is thus necessary to combine both network visualization and classical graph representation, or drop networks. The most innovative researchers in the digital humanities made the latter choice and recently turned back to this traditional visualization with time on the abscissa axis.[36] That's why Alain Berthoz, famous neurophysiologist and member of the Academy of Sciences, advocated for "simplex" solutions, *i.e.* simple solutions to complex problems.[37]

Beyond this debate over visualization, the semantic web seems even more important. Without the available linked open data platforms and the semi-automatic realignment they allow, data enrichment would have been an immensely long or impossible process and the lack of contextualization of the resulting networks would have been a strong impediment. Linked Open Data thus constitutes a desirable horizon in the research field, but also a

[35] http://www.getty.edu/research/tools/provenance/zoomify/index.html, accessed January 5, 2017.

[36] Matthew Lincoln, Abram Fox, "The Temporal Dimensions of the London Art Auction, 1780–1835", British Art Studies, Issue 4, https://doi.org/10.17658/issn.2058-5462/issue-04/afox-mlincoln.

[37] Alain Berthoz, *La simplexité* (Paris: Odile Jacob, 2009).

challenge: how can we give a total access to databases without falling into the problem of free riding?

At the other end of the chain, the gathering of the open access information requires human supervision. The threat is what Eli Pariser calls the "filter bubble"[38] (i.e. when websites' algorithms guess and select personalized searches and news streams) and the errors of Word-Sense Disambiguation[39] (i.e. when the identification of the sense of a word, which has multiple meanings, is stopped by a statistical algorithm that will judge that such alignment will be more likely to be relevant). Because of these two effects, the probability to discover less known entities is increasingly reduced. This is a danger for art history, which seeks to understand the role of art and artists at a given historical moment. Part of this task entails bringing forth forgotten figures, such as Hendrick Leys and Edmond Tschaggeny — an effort which may be compromised by filter bubbles and word-sense disambiguation. It is thus essential to encourage the publication of datasets in art history, which seem to be minimal in the era of Big Data.

[38] Eli Pariser, *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*, (London: Penguin Books, 2011).

[39] Roberto Navigli, "Word Sense Disambiguisation: A Survey". in *ACM Computing Survey (CSUR)*, 2009, vol. 41, n°2: 10.

Figure 3: Network of the artists whose artworks were sold in a same Parisian auction sale in 1868. The size of the node corresponds to the average hammer price, the color represents the artist's gender. The darker the edge, the more auction sales in common.

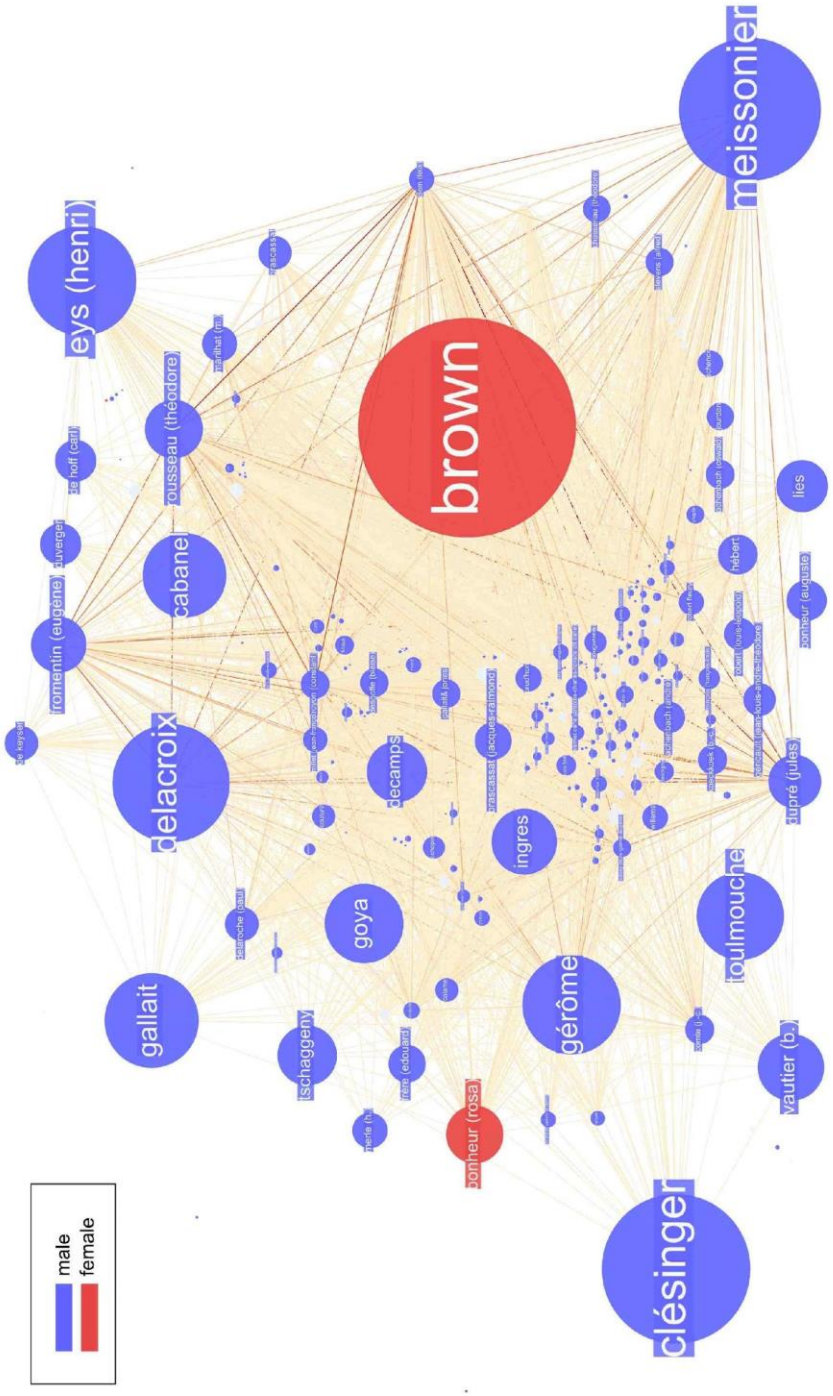

Figure 4. Network of the artists whose artworks were sold in a same Parisian auction sale in 1868. The size of the node corresponds to the average hammer price, the color represents the artist's nationality. The darker the edge, the more auction sales in common.

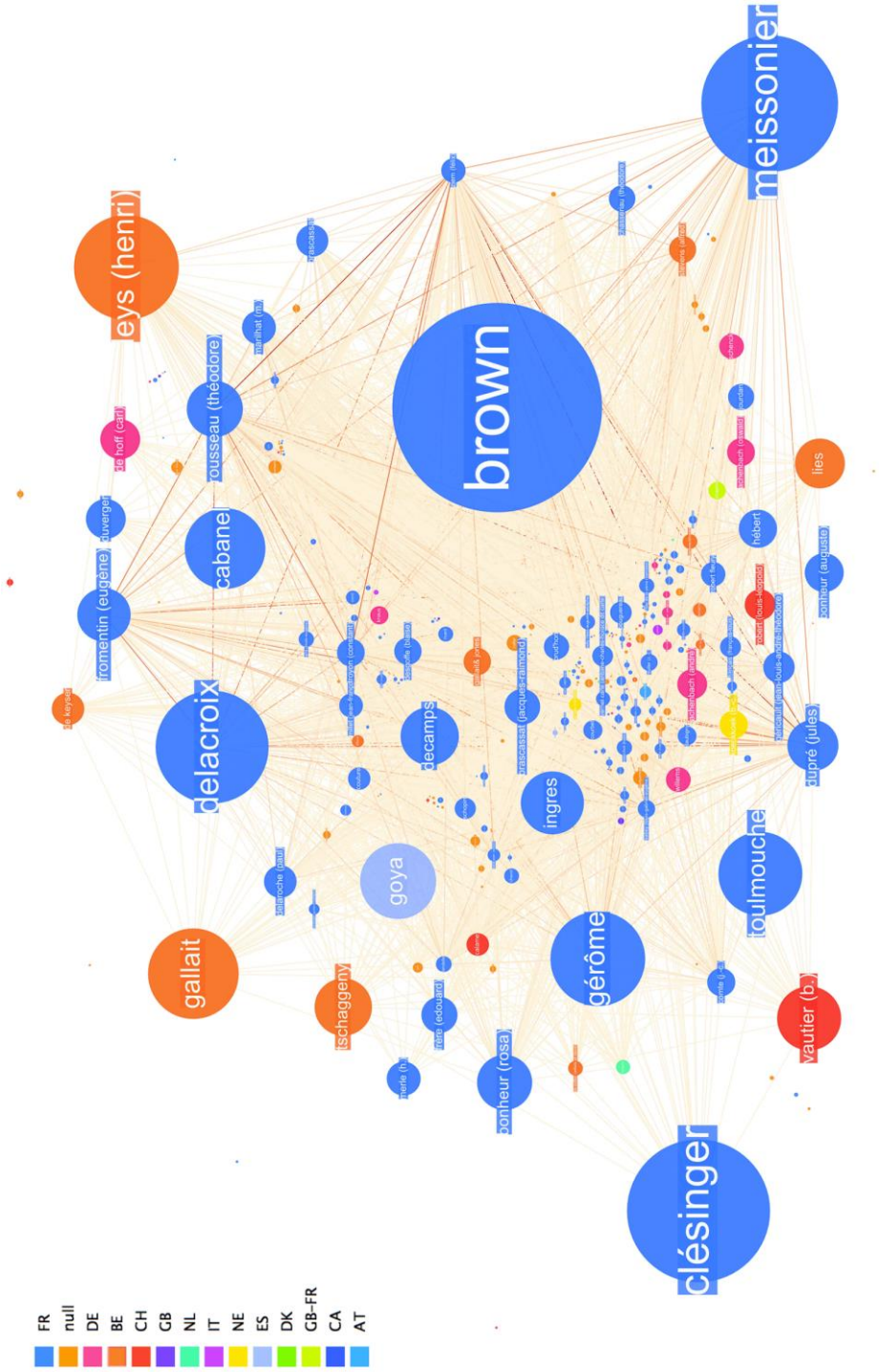

Figure 5a. Network of the artists whose artworks were purchased by a same buyer in Paris in 1868. The size of the node corresponds to the average hammer price, the color represents the artist's nationality. The darker the edge is, the more purchasers in common. The Force Atlas 2 spatialization algorithm has been chosen.

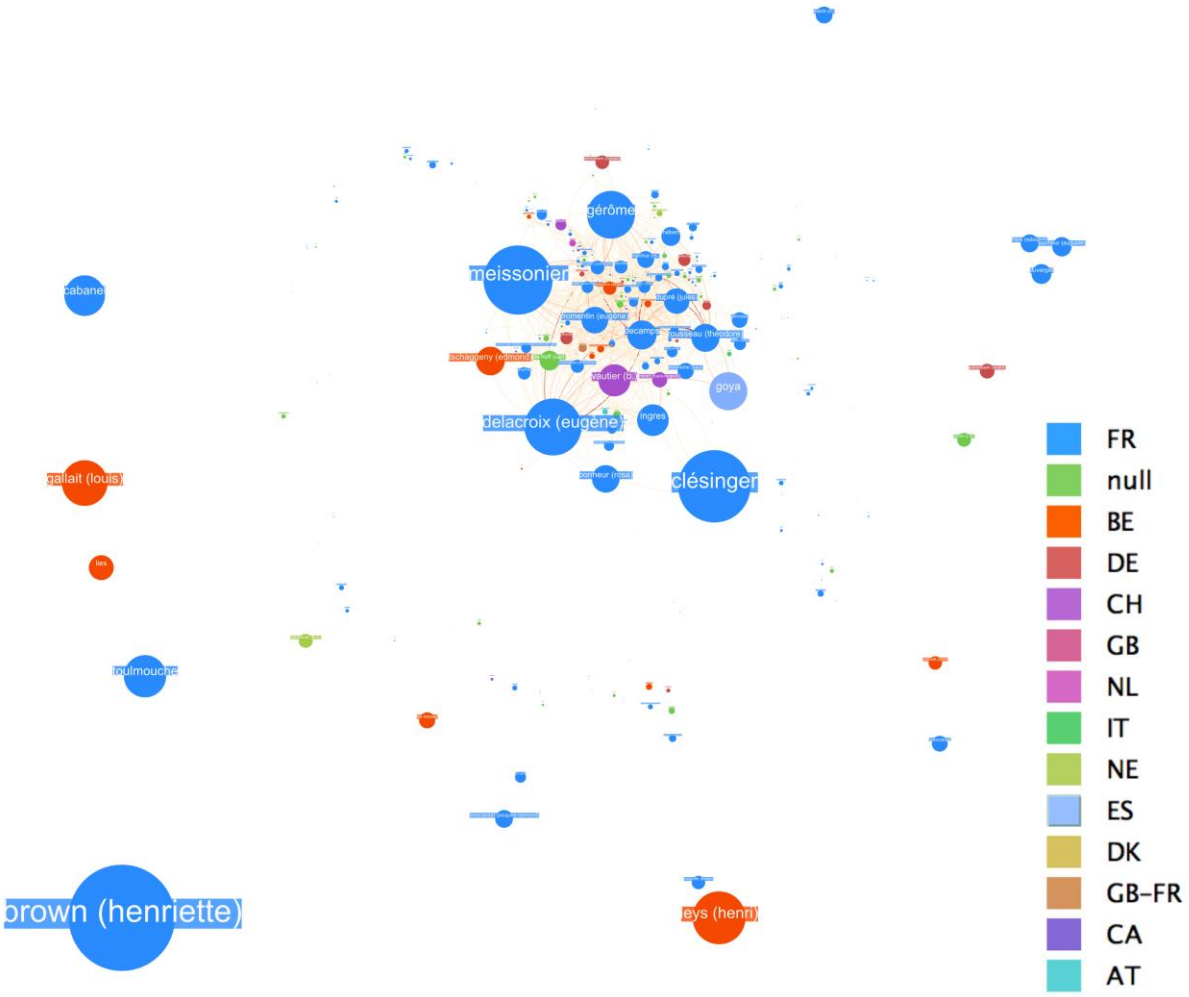

Figure 5b. detail of the core of the Fig. 5a network.

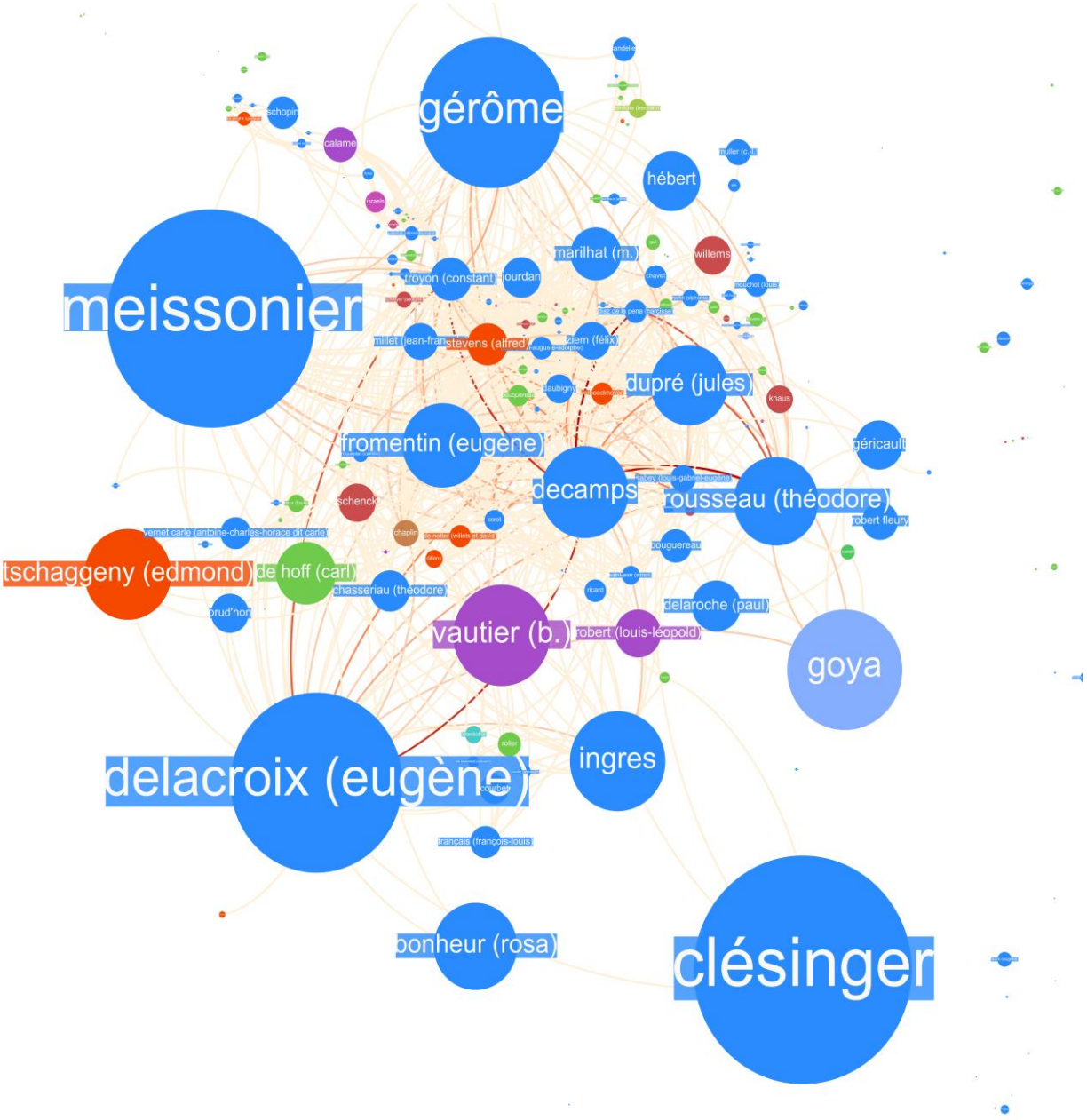

Figure 6. Network of the buyers who purchased the artworks by a same artist in Paris in 1868 (detail). The size of the node corresponds to the number of artworks that were purchased.

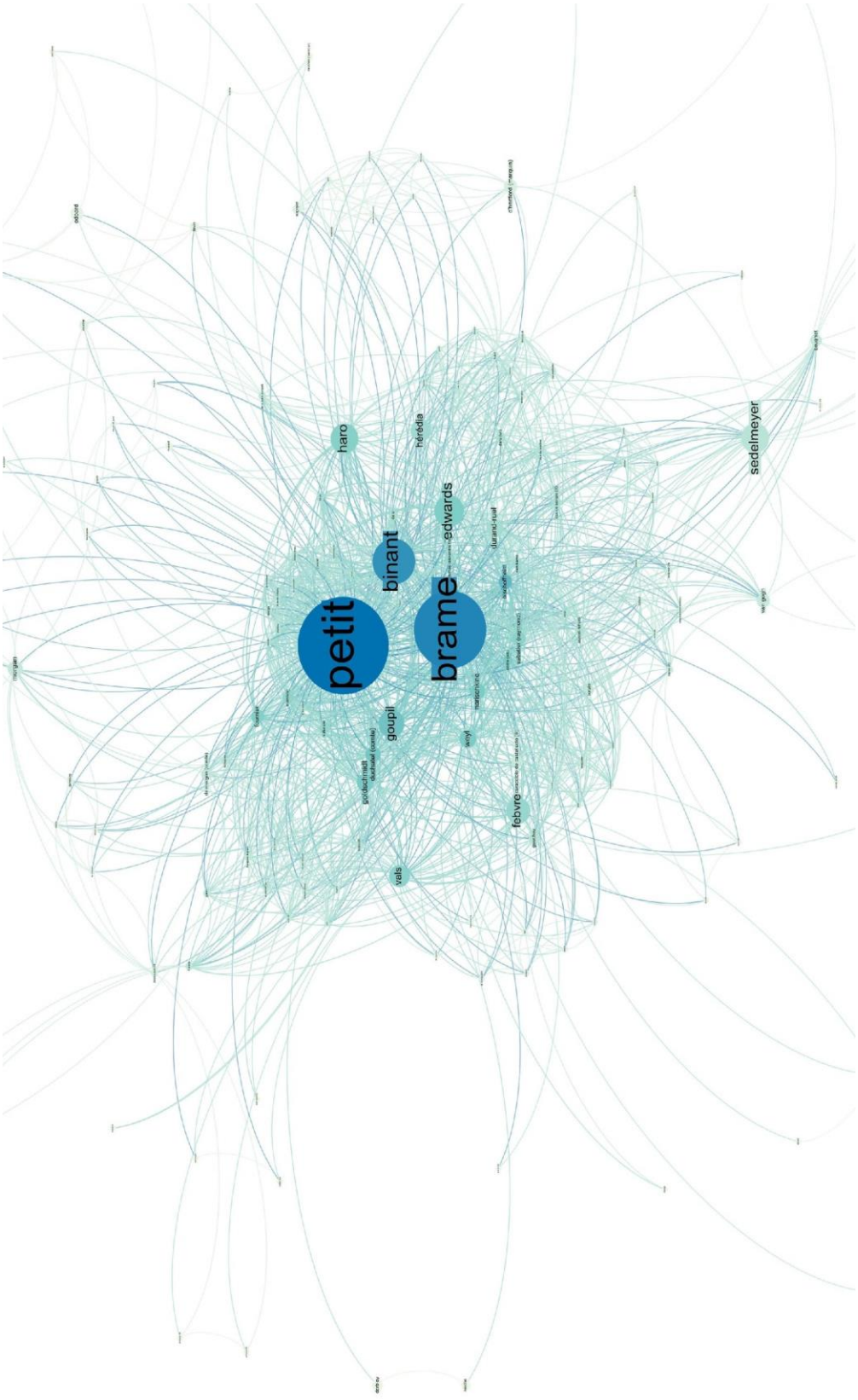

Figure 7. Comprehensive network of the buyers and the artists they purchased in common, regarding the Parisian auction sales of "modern paintings" in 1868 (detail).

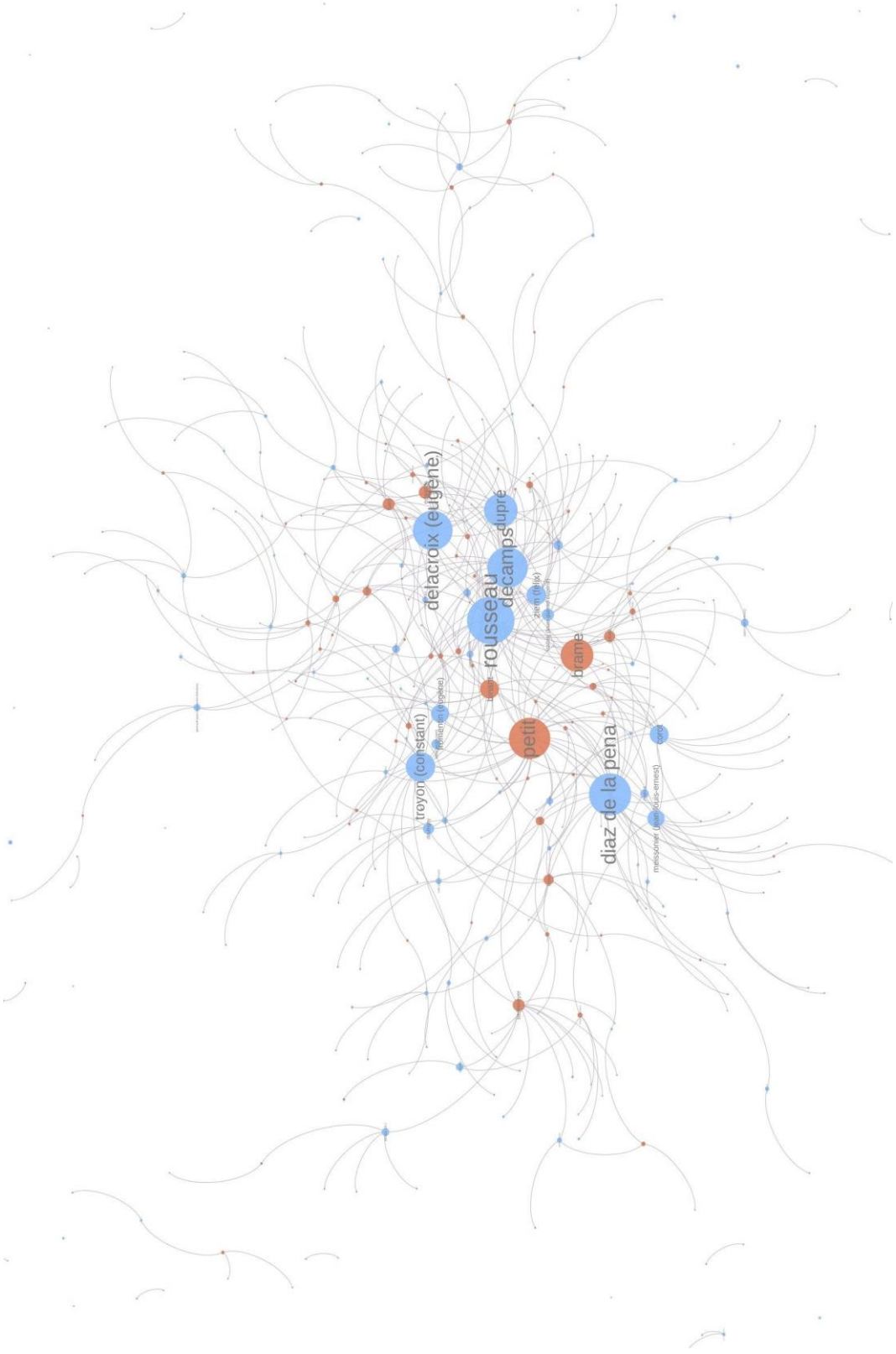

Figure 8. Network of the artists whose artworks were purchased by a same buyer in Paris in 1868. The size of the node corresponds to the average hammer price, the color represents the artist's nationality. The darker the edge is, the more purchasers in common. The Fruchterman Reingold spatialization algorithm has been chosen.