Graphs and Social Systems

Vincent Labatut, Rosa Figueiredo

JINIS

Journal of Interdisciplinary Methodologies and Issues in Science

VOLUME 2
JULY 2017

Graphs & Social Systems
Graphes & Systèmes Sociaux

GUEST EDITORS

Vincent Labatut <vincent.labatut@univ.avignon.fr>
Rosa Figueiredo <rosa.figueiredo@univ.avignon.fr>
Laboratoire Informatique d'Avignon - LIA EA 4128

cnrs
épisciences.org
UNIVERSITÉ D'AVIGNON
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE
agorantic
Culture, Patrimoines, Sociétés numériques
Fédération de Recherche

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

Introduction to the special issue on *Graphs & Social Systems*

Vincent LABATUT[*1,2] **and Rosa FIGUEIREDO**[*1,2]

[1]Laboratoire Informatique d'Avignon – LIA EA 4128, France
[2]Agorantic – FR 3621, France

*Corresponding author: {firstname.lastname}@univ-avignon.fr

DOI: 10.18713/JIMIS-300617-2-0
Submitted: 24 April 2017 - Published: July 10 2017
Volume: **2** - Year: **2017**
Issue: **Graphs & Social Systems**
Editors: *Rosa Figueiredo & Vincent Labatut*

I INTRODUCTION

The principle of the *Journal of Interdisciplinary Methodologies and Issues in Science* (JIMIS) is that each issue is a special one, dedicated to a specific topic and handled by guest editors. This issue (the second of the journal) focuses on the use of graphs (and associated analysis tools) to model and study social systems. The guest editors for this issue are Rosa Figueiredo and Vincent Labatut (cf. Section III for affiliations and other details).

A social system can be viewed as a set of entities such as persons, social groups and institutions, interconnected through various types of relationships, in such a way that the whole constitutes a coherent structure. Social system analysis is an inherently interdisciplinary academic field, which emerged from sociology, statistics, social psychology, graph theory, and other domains. For the last few decades, and in parallel with the development of the network science field, graph-based approaches dedicated to this purpose have gained a significant following in social sciences and humanities, and there are now automatic tools commonly available for end-users.

Thanks to the very generic nature of graphs, it is possible to take a method designed to handle a specific system, and apply it in a completely different context (with various levels of adjustment). For instance, a method allowing to detect functionally important proteins in a biological network can be used to identify key-players in a social network. However, due to lexical, methodological and cultural differences, being aware of the methods developed in other fields can be truly challenging for a researcher.

The goal of this special issue is to try to bridge this gap, by exposing researchers from Computer Sciences and from Humanities and Social Sciences to different tools and usages of the concept of graph, coming from out of their field. The selected articles describe graph analysis methods and models, as well as their application to specific social systems.

II IN THIS ISSUE

This issue originates from the *Seminar on Graphs & Social Systems* that took place on the 18th of March 2016 in Avignon, France[1], and that focused on the same topic: using graphs to model and study social systems. During this day of presentations, a variety of works was described to a relatively large and interdisciplinary audience. This motivated the guest editors to propose a special issue to JIMIS, aiming at presenting the same types of works.

In order to widen the audience, the call for paper was open. We initially received a total of 9 answers, 4 of which went all the way through the editorial process. This includes two rounds of reviews by at least 3 reviewers representing at least two disciplines, in order to give the authors both methodological and application-related feedbacks. There was an additional invited contribution (Patrick Doreian's article), which underwent the same process. Overall, the 5 articles of this special issue cover a large scientific range, from theoretical to applied aspects, fitting the scope of the journal.

In his article *Reflections on Studying Signed Networks*, Patrick Doreian (2017) takes advantage of his extensive experience of signed graphs to propose a review of the works related to structural balance. In a signed graph, each link is associated to either a positive or a negative sign, which allows to model antagonistic relationships. This makes this type of graphs particularly appropriate to represent polarized social systems. *Structural Balance* is a theory proposed by Heider (1946) to explain the distribution of signs in such networks. Doreian reviews the main methods and results related to structural balance, but he also identifies limitations and open problems. In particular, he identifies the separation between substance and methods as a major issue, i.e. developing data analysis methods independently from social science theories, and defining such theories without any regards for empirical results. This observation particularly fits the objective of this issue, since it is intrinsically interdisciplinary.

Alain Guénoche (2017), in his article entitled *Analyse des Préférences et Tournois Pondérés* (in French), assumes the existence of multiple rankings on n items and studies two ranking problems. The first problem consists in establishing a total order on a set of n items, the second one is the selection of the k best elements among the n items. Both problems reduce to minimizing the number of preferences that go against the individual choices and appears in experimental studies as procedures for vote counting. The developed methods are based on the representation of majority preferences as a complete oriented graph. Although the subject has been treated many times in the literature, there is a new contribution here, taking the form of an optimal selection method in the case where n is sufficiently small.

In their work *Brazilian Congress structural balance analysis*, Mario Levorato & Yuri Frota (2017) study the behavior of Brazilian politicians in the period between 2011 and 2016. Inspired by a previous work (Mendonça *et al.*, 2015), they extract and analyze a collection of signed networks representing voting sessions of the lower house of the Brazilian National Congress. The solutions obtained by solving Correlation Clustering problems on the extracted signed networks are the basis for investigating deputies voting networks, as well as questions about loyalty, leadership, coalitions, political crisis, and social phenomena such as polarization. Their work contributes to filling the gap identified by Patrick Doreian (2017) between substance and methods in the application of the structural balance theory.

Jérôme Kunegis, Fariba Karimi & Sun Jun (2017) propose a model to explain the mechanism of preferential attachment, in their article *The Problem of Action at a Distance in Networks and*

[1] https://jgss.sciencesconf.org/

Volume: **2** - Year: **2017**, *DOI: 10.18713/JIMIS-300617-2-0*

the Emergence of Preferential Attachment from Triadic Closure. In the context of complex networks, the concept of *Preferential Attachment* was introduced by Barabási and Albert (1999) to explain the scale-free property in real-world networks (power law-distributed degree). It states that when a new node appears in an existing network, it tends to get attached to nodes already possessing many connections. The thesis defended by Kunegis *et al*. is that there actually is a lower level process behind preferential attachment. They argue that it is the consequence of the joint occurrence of two events: first, the considered network is only partially known (some nodes are hidden), and second, new links are formed by triadic closure in this underlying network, giving the impression of a non-local process. They formalize their idea as a model and show its validity both analytically and empirically. Although this sounds very mathematical, we encourage people from all fields to read this article, because its authors managed to introduce its core concepts and discuss its main results in a very intuitive way, putting every idea in perspective with concrete, illustrative observations.

The article *Analyse de rśeaux criminels de traite des êtres humains: modélisation, manipulation et visualisation* (in French) by Bénédicte Lavaud-Legendre, Cécile Plessard, Guy Melançon, Antoine Laumond & Bruno Pinaud (2017) is very illustrative of what an interdisciplinary work can be. It gathers specialists from 3 distinct fields: Law, Sociology and Computer Science, whose goal is to identify and study prostitution networks. The article describes the methods used to build a database from a large quantity of very raw data, and to extract various types of social networks from this database. The raw material is constituted of all the investigation material gathered by the police when working on related cases, including: interviews, transcriptions of phone tapping, seized documents, surveillance reports, and so on. This extreme heterogeneity makes the network extraction task very difficult, and the authors describe how they tackled this challenge in an incremental way. They also give the first results obtained by studying the obtained networks: in particular, they identify a limited number of operational roles, such as prostitute, sponsor, recruiter, etc. An important characteristic of this work is its applied nature: it aims at providing a platform which will be used by the French *Police Judiciaire* (judicial police).

III SCIENTIFIC COMMITTEE

Each issue of JIMIS deals with a special topic, and as such it has its own scientific committee. Here are the members of the scientific committee selected for this *Graphs & Social Systems* issue (in alphabetical order):

Vincent Labatut
Complex networks analysis, Information retrieval
Université d'Avignon et des Pays de Vaucluse / EA 4128 Laboratoire Informatique d'Avignon (LIA) / FR 3621 Agorantic, Avignon (France)

Guillaume Marrel
Politics and history, State politics and policy, Representation and electoral systems
Université d'Avignon et des Pays de Vaucluse / EA 3788 Laboratoire Biens, Normes, Contrats (LBNC) / FR 3621 Agorantic, Avignon (France)

Pierre-Henri Morand
Game theory and social networks, Public economics
Université d'Avignon et des Pays de Vaucluse / EA 3788 Laboratoire Biens, Normes, Contrats (LBNC) / FR 3621 Agorantic, Avignon (France)

Michael Poss
Integer programming, Robust optimization, Network design
UMR 5506 Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Montpellier (France)

Cristina Requejo
Combinatorial optimization, Integer linear programming, Network design
Center for Research & Development in Mathematics and Applications (CIDMA)
Department of Mathematics, University of Aveiro, Aveiro (Portugal)

David Savourey
Integer Linear Programming, Scheduling
UMR 7253 Heuristique et Diagnostic des Systèmes Complexes (HeuDiaSyC), Compiègne (France)

References

Barabási A.-L., Albert R. (1999). Emergence of scaling in random networks. *Science 286*(5439), 509. doi:10.1126/science.286.5439.509.

Doreian P. (2017). Reflections on studying signed networks. *Journal of Interdisciplinary Methodologies and Issues in Science 2*, 2.1–2.14. doi:10.18713/JIMIS-170117-2-1.

Guénoche A. (2017). Analyse des préférences et tournois pondérés. *Journal of Interdisciplinary Methodologies and Issues in Science 2*, 3.1–3.16. doi:10.18713/JIMIS-170117-2-2.

Heider F. (1946). Attitudes and cognitive organization. *Journal of Psychology 21*(1), 107–112. doi:10.1080/00223980.1946.9917275.

Kunegis J., Karimi F., Sun J. (2017). The problem of action at a distance in networks and the emergence of preferential attachment from triadic closure. *Journal of Interdisciplinary Methodologies and Issues in Science 2*, 5.1–5.13. doi:10.18713/JIMIS-140417-2-4.

Lavaud-Legendre B., Plessard C., Melançon G., Laumond A., Pinaud B. (2017). Analyse de réseaux criminels de traite des êtres humains : modélisation, manipulation et visualisation. *Journal of Interdisciplinary Methodologies and Issues in Science 2*, 6.1–6.25. doi:10.18713/JIMIS-170117-2-5.

Levorato M., Frota Y. (2017). Brazilian congress structural balance analysis. *Journal of Interdisciplinary Methodologies and Issues in Science 2*, 4.1–4.27. doi:10.18713/JIMIS-280217-2-3.

Mendonça I., Figueiredo R., Labatut V., Michelon P. (2015). Relevance of negative links in graph partitioning: A case study using votes from the European Parliament. In *2nd European Network Intelligence Conference*, pp. 122–129. doi:10.1109/ENIC.2015.25.

A ACKNOWLEDGMENT

This special issue was started during the *Seminar on Graphs and Social Systems* (JGSS – Journée Graphes et Systèmes Sociaux) which took place in Avignon (France), on the 18th of March 2016. This event was supported by the *FMJH Program Gaspard Monge in Optimization and Operations Research*, *EDF* on the project N^{o}2015-2842H, the *Agorantic* research federation FR 3621, and the *Laboratoire Informatique d'Avignon* EA 4128.

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

Reflections on Studying Signed Networks

Patrick DOREIAN[*1,2]

[1] Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
[2] Department of Sociology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

*Corresponding author: pitpat@pitt.edu

DOI: 10.18713/JIMIS-170117-2-1
Submitted: June132016 - Published: January172017
Volume: **2** - Year: **2017**
Issue: **Graphs & Social Systems**
Editors:*Rosa Figueiredo &Vincent Labatut*

Abstract

Despite considerable success, the balance theoretic approach to studying signed relations has encountered some serious problems, both substantive and methodological. The more consequential problems are outlined along with arguments for why solving them is so critical. An agenda of research problems is laid out with many juicy problems to solve. These reflections, while setting a context in prior work, are far more concerned about looking to the future and identifying problems whose solutions hold the potential for transforming the field.

Keywords

Signed networks;Structural balance;Relaxed structural balance;Multiple processes;Open problems.

I was surprised by, and most appreciative of, the invitation to reflect on studying signed networks for this special issue. While there are multiple approaches to examining such networks, the tack taken here deals with one line of work with which I am most familiar. These reflections have two broad components. One is to provide a partial summary of this fruitful line of research. More importantly, the second component outlines some difficult problems that have emerged for which solutions must be found. While it was tempting to confine attention to some of the successes of this approach, these serious problems raise major questions. Further, if these problems are solved, other avenues of inquiry will be opened. Future challenges are more consequential than past successes. So, this reflection can be read also as an invitation to join the quest.

I STRUCTURAL BALANCE THEORY

For too long, social network researchers confined their attention largely to networks featuring only positive ties. Yet, human relations have been known for a long time to be signed. An

important early paper on signed relations is the statement by Heider (1946) which was foundational for the line of research considered here (Taylor, 1970). As with most profound statements, it was a simple one. Its essential elements are shown in Figure 1. Also, it was substantively based, a critical feature. Indeed, there is far more to doing social network research than the analysis of network data (Robins, 2015). From a social science perspective, there has been an unfortunate separation of substance and methods within some social sciences. A substantive account of social phenomena is a logically consistent account attempting to explain the phenomena using theoretical ideas. Methods can be viewed as a set of techniques designed to analyze data to test theories. When theories are developed with little or no regard for data or when methods are developed as ends in and of themselves, this disjunction is problematic. Ideally, the two need to be integrated in a coherent fashion. See Bruscoet *al*. (2011) for a fuller account of this distinction.

Heider's idea came from thinking about the cognitive images three actors, denoted by p, o and q, could have of their signed relations. Positive ties are represented by solid lines in Figure 1. The dashed ties represent negative ties. According to Heider, the triples in the top row are 'comfortable' for p, o and q. Reading across the top row of Figure 1, the triples have been captured by four folk statements: i) a friend of a friend is a friend; ii) an enemy of a friend is an enemy; iii) a friend of an enemy is an enemy; and iv) an enemy of an enemy is a friend. Put differently, they are balanced. In contrast, all the triples in the bottom row were considered as discomforting. Actors in these triples were thought to be motivated to reduce this discomfort by changing one sign in the triple. Another way of representing the signs is numerical by using 1 (for positive ties) and -1 (for negative ties). The sign of a triple is the product of the signs in it. When the product is 1, the triple is balanced. It is imbalanced if the product is -1. Figure 1 is organized to separate the balanced and imbalanced triples into two panels.

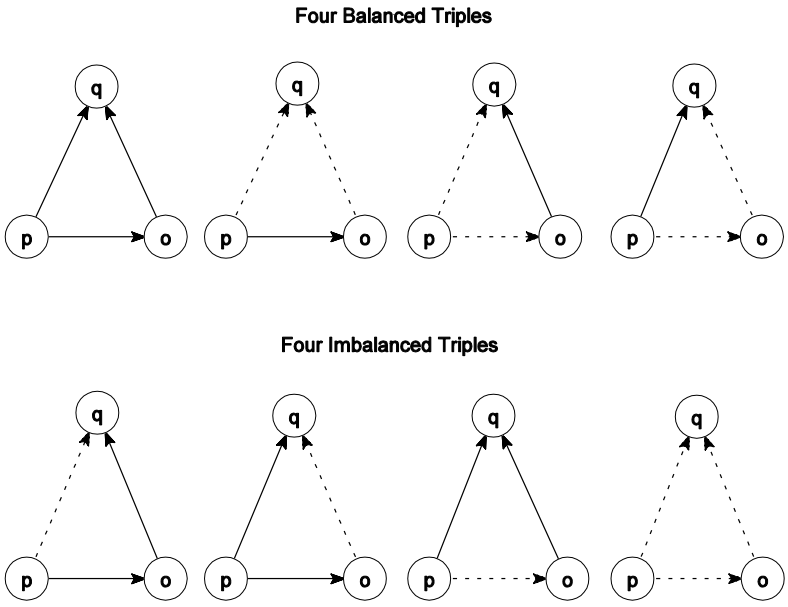


Figure 1: Heider's balanced and imbalanced triples.Note: Solid lines represent positive ties and dashed lines represent negative ties.

Considering triples in this fashion can be extended to networks by computing the signs of all of the triples in a network. A balanced network has only balanced triples while any network with at least one imbalanced triple is imbalanced. The larger the number of imbalanced triples, the greater the imbalance of the signed network. Two strands of thoughts followed this extension.

The first was methodological, in which the task was to compute a measure for the (extent of) imbalance of a network. An early measure of the imbalance of a signed network is the proportion of imbalanced signed triples among all signed triples (Hummon & Fararo, 1995). Computing this for complete signed networks is unproblematic. But when these networks are not complete, the measure of imbalance can be recast as the proportion of balanced cycles of any length. Computing such a measure turned out to be a difficult computational problem. The second strand was substantive. It took the form a seemingly plausible empirical proposition: Signed networks move towards a balanced state.

Cartwright & Harary (1956) picked up on this idea for signed networks, albeit for networks outside the minds of actors. They formulated and proved a theorem: When a complete signed network is balanced, the vertices representing actors can be partitioned into two subsets of actors with a striking property: All the ties between actors within a subset are positive while all of the negative ties are between actors in different subsets. Davis (1967) noted some human groups split into more than two mutually hostile subsets and suggested the all negative triple (bottom right in Figure 1) be considered as balanced. Doing this led to another theorem where there could be more than two subsets of actors with the same property. Doreian & Mrvar (1996) labeled them 'the structure theorems' of balance theory. The results were extended easily to signed networks that were not complete. Letting k denote the number of clusters (called positions), $k = 2$ is for the Cartwright and Harary theorem and $k > 2$ applies for the Davis theorem.

A completely different strand of research for studying social networks takes the form of blockmodeling, where the goal is to take a large (and potentially complex) network and partition the vertices and the ties to create a simpler and far smaller 'image' (blockmodel) of the network. The early uses of blockmodeling were instrumental in developing a unique approach to discerning the underlying structure of networks with a concern for structural roles. Most were based on structural equivalence (Lorrain & White, 1971) with two algorithms coming to dominate the empirical analyses. They were due to Breiger *et al.* (1975) and Burt (1976). Batagelj *et al.* (1992a, b) and Doreian *et al.* (2005) provided an alternative conception of how to partition social network while remaining faithful to the blockmodeling enterprise.

Doreian & Mrvar (1996) noticed that the balance structure theorems permitted a clear connection between studying signed networks and blockmodeling. They defined a positive block as having only positive (and null) ties while a negative block has only negative (and null) ties. The structure theorems imply that positive blocks are located only on the main diagonal of the blockmodel and negative blocks are all off the diagonal. They proposed a fast heuristic algorithm for partitioning signed networks based on structural balance. In principle, this can be applied to any signed network. However, partitioning large networks is problematic given this being an NP-hard problem.

The Doreian & Mrvar (1996) procedure used an alternative measure of imbalance taking the form of the number of ties whose sign change (or removal) created a balanced network. This is the line index introduced by Harary *et al.* (1965). More formally, the criterion function for a binary network (where the ties are $+1$, 0 or -1) is the total number of positive inconsistencies (positive ties in what is thought to be a negative block), denoted by P, and the total number of negative inconsistencies (negative ties in positive blocks) denoted by N. A general measure of how poorly a blockmodel fits the data is given by a criterion function, C_f, where $C_f = \alpha N + (1 - \alpha)P$, where $0 < \alpha < 1$. With $\alpha = 0.5$, the two types of inconsistencies are weighted equally, a convention that has been followed consistently. The algorithm was implemented in Pajek (Batagelj & Mrvar, 1998).

Everything looked good both for partitioning signed networks and computing the balance of signed networks. Indeed, many useful analyses were done with substantive implications. But

empirical reality was not so kind for these endeavors because substantive and methodological problems were exposed. The substantive ideas were not supported by empirical data and unproductive ideas were pursued despite these problems. Also, the methods used were shown to have major problems. An examination of them follows in the next section.

II MAJOR PROBLEMS WITH BALANCE THEORY

The problems start with the basic empirical hypothesis informing the approach for a long time. Is there really a universal tendency towards balance? The evidence, what little there is, suggests not. Doreian & Krackhardt (2001) examined the well-known temporal Newcomb (1961) social network data (documented by Nordlie, 1958). If the simple empirical hypothesis regarding a universal movement towards balance was correct, then the number of all of the balanced triples of Figure 1 would increase through time while all of the negative triples would decrease. This idea was shattered. Overall, two of the balanced triples became less frequent over time while two of the imbalanced triples became more frequent! While four triples did behave as expected under the empirical hypothesis, the behavior of the other four was bad news.

Hummon & Doreian (2003) conducted a simulation of 'Heider actors' behaving in ways fully consistent with Heider's ideas. Their results were even more devastating. First, many networks did *not* end in a balanced state. The reason for this was quite simple. Attempts to create balance in one triple could – and frequently did – have impacts in other triples. Creating balance in one triple can imply imbalance in one or more other triples. Achieving balance in signed networks is not simple. Second, the sequences of network structures, as the ties evolved over time, were very long with both increases and decreases of imbalance. The implications were obvious: changes in the overall balance of signed networks are not unidirectional. Most certainly, they are *not always towards* balanced configurations. This was an important, albeit largely overlooked, result. See also Robins & Kashima (2008). The basic hypothesis of networks moving towards balance is false. The almost exclusive focus on this hypothesis had unfortunate consequences as described in Section 3.2.

In the context of human signed relations, Doreian & Mrvar (2009) asked another question: Is the blockmodel structure implied by structural balance overly restrictive and counter-productive? They answered in the affirmative. The statement of the structure theorems (described above) implied a specific block-structure with positive blocks on the main diagonal and negative blocks of the diagonal of the blockmodel. The assumption was that structural balance was the only process operating. This changes if other processes are in play. Differential popularity is clearly present in human groups, as is differential unpopularity. The former clearly points to positive blocks *off* the main diagonal of the blockmodel. If they exist and, perhaps, also negative blocks on it, how could signed blockmodeling accommodate this? Asking this question led them to propose the notion of 'relaxed structural balance' under which positive and negative blocks could appear anywhere in the blockmodel image. The criterion function for imbalance was not changed. Note this algorithmic change was driven by substantive concerns. For the human networks, they studied, much better fits to the data were achieved. See Doreian & Mrvar (2014) for a more extended study of signed networks using relaxed structural balance. There were positive off-diagonal blocks. And there was a 'den of vipers' creating a negative block on the main diagonal of the blockmodel image. If the structural features identifiable using relaxed structural balance are present in a signed network, this helps explain why classical balance theory failed so often to account for the structural changes of signed networks over time.

Signed relations can occur in many contexts. The core idea in structural balance can be reformulated by using a more general notion of *consistency* among a set of ties within a signed network. Of course, the notion of consistency must be specified in ways that will depend on

contexts. As an example, consider the phenomenon of the US Supreme Court overturning earlier decisions. The positive ties are later decisions citing earlier decisions as legitimate precedents. Negative ties are for later decisions overturning earlier decisions completely. Figure 2 shows two triples of decisions. In the top triple Decision 3 overturned Decision 2 which had affirmed decision 1. Yet Decision 3 also affirmed Decision 1. In the second triple, Decision 6 affirmed Decision 5 which also affirmed Decision 4. Yet Decision 5 overturned Decision 4. Clearly, the social psychological presumptions of balance theory do not apply in this context. But the triples seem fundamentally inconsistent in the sense of being contradictory. In contrast, if the two triples each had only positive ties, there would be consistency for the decisions.

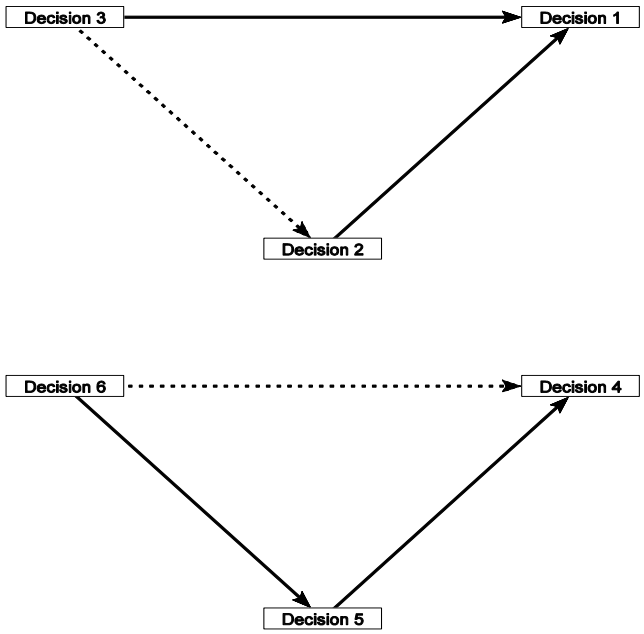


Figure 2: Examples of inconsistent triples for the US Supreme Court.Note: Solid lines are affirmations and dashed lines are overturning pairs.

Thinking in terms of signed consistency permits a natural extension beyond interpersonal signed networks. In this spirit, Doreian & Mrvar (2015) studied the structure of the networks of international relations over time using the Correlates of War (CoW) data. (See Pevehouse *et al.* (2004) for details regarding these data.) The definitions of the signed relations they studied are simple to state. First, positive ties between nations are defined by joint memberships in alliances, being in unions of states and sharing inter-governmental agreements. Second, negative ties are for two states being at war, being involved in border disputes, conflicting with each other without military involvement, or having sharp ideological or policy disagreements. This creates a signed network for nations as the actors. Their study featured blockmodeling and measuring imbalance in the overall network[1].

On tracking balance over time in the international system, Doreian & Mrvar (2015) showed there was no consistent movement towards balance over time. This is revealed clearly in Figure 3 where the amount of imbalance is plotted over time. The underlying assumption was that there would be signed consistency in signed relations but without appealing to social psychological principles. The ideas underlying structural balance were extended to other

[1]Also, when there is a negative tie between states otherwise having a positive tie, the negative tie is used.

networks as an organizing framework in terms of consistency. We were not alone in formulating this extension. See also Mendonça*et al*. (2015) and Vinogradova & Galam (2014). Such approaches can be developed in different fields.

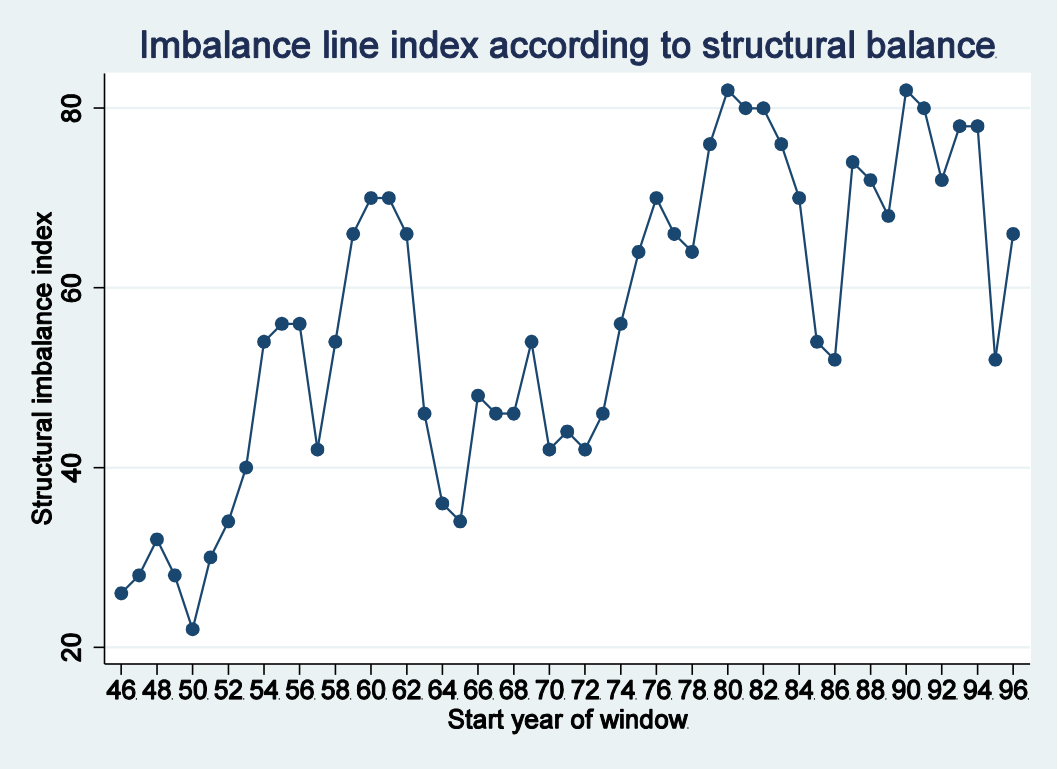


Figure 3: Imbalance measure over time in the international system, 1945-1999.

Establishing the block structures of the signed international relations revealed two serious problems. The first was that the Doreian-Mrvar balance partitioning algorithm failed badly in the way it handled the positive ties. The definition of positive blocks was the problem leading to larger clusters than was reasonable. The second was the failure of the Traag&Bruggeman (2009) community detection algorithm in handling negative ties. Here the problem was the use of modularity which works well for positive ties. But it failed for negative ties by creating a far too fragmented block structure.

Two-mode networks also can be signed. One such network takes the form of Justices in the US Supreme Court (forming one set of objects) agreeing/concurring with or dissenting from decisions made by this court, (the second set of objects). Another example has the voting behavior of nations casting votes in the United Nations General Assembly (UNGA) for or against resolutions. Both involve another construction of what constitutes a signed tie as positive votes for items and negative votes against the same items. Mrvar & Doreian (2009) established an algorithm for partitioning signed two-mode networks, also included in Pajek. An example of a partitioned Supreme Court signed network is shown in Figure 4. The black squares represent votes for decisions, the red squares are for dissenting (negative) votes and white squares are for when justices took no part in a decision. The Justices are partitioned into the assumed conservative and liberal wings of this court for its 1995 term. The decisions, labeled by the substance of the issue involved, are partitioned per the distinct *patterns* of how the justices voted. However, this partition was *not* established with the two-mode partitioning algorithm for it failed completely, especially for clustering the decisions. Indeed, a close look at the pattern of votes suggest it was doomed from the start. The empirical partition structure

was too nuanced and complex. Instead, the partition was established through a close inspection of the data.

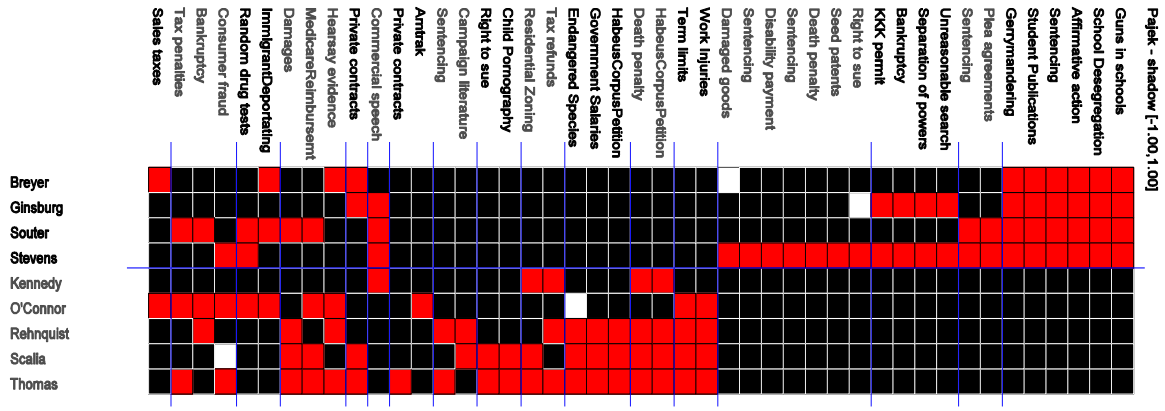


Figure 4: Supreme Court voting for the 1994-1995 term, non-unanimous decisions.Note: Black squares represent votes for a decision, red squares, votes against a decision and white squares for taking no part.

In a separate study, Doreian *et al.* (2013) presented partitions of UNGA voting on resolutions. For this larger signed network, the two-mode algorithm was successful – but the delineation was established only with great difficulty. Clearly the operation of the two-mode partitioning algorithm merits closer attention.

All the problems described in this section exposed some limitations with the general balance theoretic approach, both substantively and methodologically. These and some additional open problems are discussed in the next section. The current signed partitioning 'state of the art' cannot be accepted as final. Of course, this holds for any method despite proponents of 'their' methods appearing to think enough has been done. More importantly, in this context, the expression 'state of the art' needs to be changed to 'states of the art' implying a need for greater communication between the states. Indeed, Arefet *al.* (2016) provide an exact method for computing for the frustration index (level of imbalance) in signed networks which appears to outperform the results reported in Doreian & Mrvar (2014).

III SOME CONSEQUENTIAL ISSUES AND OPEN PROBLEMS

The issues identified in Section 2 must be tackled. Moreover, both substantive and methodological developments are necessary. They need to be coupled. The days of single 'cookie-cutter' approaches to data analysis are long gone – or should be.

3.1 Substantive foundations for balance theory

One presumption of balance theorists was that balance was a singular process. At best, given the results of Doreian & Krackhardt (2001), it is a set of processes. Some of them might work while other do not. Discerning which do and which do not requires carefully collected data. Abell & Ludwig (2009) posit the notion that some actors are more tolerant of imbalance than others. They studied balance processes through simulations so their results remain suggestive only regarding empirical reality. Examining the merits of this idea requires detailed and careful measurement in field settings. While differential tolerance is a useful substantive consideration it also points to the inclusion of actor attributes. Classic structural balance is limited, perhaps even doomed to failure, if such attributes are ignored completely.

The key implication of the structure theorems can be viewed as being rather bleak concerning inter-group conflict. If a group splits into any number of mutually hostile subgroups, then there will be little chance for the resolution of these conflicts even if such entrenched conflicts are

damaging for collective outcomes absent conscious actions to resolve them. It would leave no room for mediation. Doreian & Mrvar (2014) suggested some actors could be motivated to mediate disputes to resolve them. But this also is no more than a speculation. Empirical information would be required to assess whether it is present in mediated conflicts. It could be that such mediators have greater tolerance for imbalance. Most likely, multiple balance processes, far more complex than envisioned by balance theorists, plus the operation of differential popularity and unpopularity, mediation and other mechanisms are likely to be in play in social settings. Understanding the operation of *multiple* processes *in conjunction* with each other necessarily requires far more complicated theories about the operation of the mechanisms involved in generating signed networks. It will mean also examining and testing *rival* hypotheses. An example of doing so is found in Doreian & Mrvar (2014) in which classical structural balance did not fare well. Alas, the data they used were limited. Much better data are needed for future studies.

Even with a focus on structural properties of networks, where attention is restricted to the network ties, there is a prior issue to consider. Kalish & Robins (2006) studied the psychological predispositions of actors in forming network ties. See also Breiger and Ennis (1979). This goes well beyond the idea of tolerance for imbalance - but is of the same ilk. The formation of network ties depends on: the extant distribution of ties; actor attributes and the *combination* of attributes for *pairs* of actors. It depends also on prior network events involving those actors (Doreian, 2002). This requires far more sophisticated theories about the generative processes, or mechanisms, of signed networks. It is another urgent but difficult problem.

3.2 On the failure of the basic empirical hypothesis of structural balance

As noted above, the basic hypothesis that human signed networks move towards a balanced state is not supported. At best, the evidence is mixed - but most contradicts it. There have been, and continue to be, attempts to assess the 'balance hypothesis' using only positive network ties. All such efforts *must* be dismissed because the dynamics of signed and unsigned networks differ greatly. Such attempts have persisted, especially within US sociology. But this extends more generally. The dynamics of signed relations involve both positive and negative ties: They are equally important. Doreian *et al.* (2015,p.312-3) showed how the separate analyses of positive and negative ties for a social group was useless. Even allowing negative ties in a secondary role to an examination of positive ties (see Esmailian*et al.*,2014) is problematic when studying *signed* relations, a view shared by Mendonça *et al.* (2015).

Even worse, the acceptance of this balance hypothesis as real, with narrowly focused subsequent attempts designed to support it, meant a far more important question was not asked. It is simple to state: *What are the conditions under which signed networks move towards balance and what are the conditions when they move away from it?* As it has been ignored for far too long, the time to address it seriously has come. The trajectories of the signed triples presented by Doreian & Krackhardt (2001) provided one prod to doing this. They suggested that balance *might* work when the tie from *p* to *o* was positive but *not* when it was negative. But this line of thought focused only on the network ties. The movement of the imbalance measure over time shown in Figure 3 also demands that this question be addressed. Of course, signed ties in international relations are very different from signed ties in human social groups. No doubt, the mechanisms involved will differ. But the need to focus on the *contexts* of relations is common to both. A special issue of *Social Networks* (volume 34, issue 1, 2012), edited by adams, Faust and Lovasi, was devoted to 'capturing context' (adams *et al.*,2012). It explored contexts for social networks from a variety of vantage points. Empirical settings are far more than the places where the data happened to be collected. This also becomes part of tackling a much deeper substantive problem.

3.3Using substance to inform blockmodeling

The practices in the early uses of blockmodeling, while important for the social networks field, had an unfortunate consequence. Algorithms were switched on, partitions were delineated, these partitions were accepted and then interpreted. This approach to data analysis is a *remarkable expression of ignorance*. Most often, network analysts know far more about the networks they study either substantively or empirically. It seems rather foolish *not* to use this knowledge. Doreian *et al.* (2005) proposed the idea of pre-specified blockmodeling as a way of incorporating knowledge. This was facilitated by two mathematical theorems. Batagelj *et al.* (1992a, b) proved that the most used types of equivalences, structural and regular, implied specific block types. If an unsigned network could be partitioned exactly in terms of structural equivalence, there were only two possible types of blocks – complete and null. For regular equivalence, they were null and one-covered (each row and each column in a block has at least one 1). They also expanded greatly the number of block types for inclusion in a blockmodeling analysis. Additionally, instead of using correlational and distance measures constructed from the network ties, blockmodels could be fitted directly by focusing on types of blocks and the distribution of ties in them. They 'reversed' the approach to using equivalence by defining new block types to create new types of equivalences. The positive and null blocks defined by Doreian & Mrvar (1996) were two new block types. In general, as new network domains are studied, the need for expanding the number of useful block types will grow for substantive reasons.

Pre-specified blockmodeling takes the form of identifying some - or all if we wish to be very ambitious - block types by location in a blockmodel. The structure theorems implicitly pre-specified the form of the blockmodel for signed networks. For a limited time, it worked well for the early collected signed networks. Pre-specification was used more successfully when using relaxed structural balance by Doreian & Mrvar (2014, 2015) as noted by Esmailian *et al.* (2014). Prota & Doreian (2016) used pre-specification to great effect for delineating the structure of rice trading (unsigned) networks in Vietnam. Despite these successes, pre-specification can be a double-edged sword, one to be used judiciously.

If substantive knowledge alone is used, there is no problem. Pre-specifying a blockmodel and fitting it to network data permits a test of the underlying substantive ideas. If the blockmodel fits there is some confirmation for these ideas. But if it does not fit, these ideas become suspect – assuming the data are good. But if empirical knowledge of the specific studied network is used too much, then there is the risk of 'over-fitting' a blockmodel and reporting only what one sees in the data. For this reason, the partition shown in Figure 4 tested nothing. While it is only an empirical description, it will be useful for formulating ideas about the temporal voting patterns of the Justices. Finding the right balance when using pre-specification will be difficult – but it must be identified.

3.4Data issues

When structural balance was first proposed, and for several decades thereafter, the default mode for collecting social network data in human groups was a fixed choice design. This design restricts the number of responses to a small number of ties, most often 3. This was expanded to allowing 'up to' some specified number. While this is a modest improvement, the use of the fixed choice design remains very suspect from a measurement perspective despite its supposed convenience. Wasserman & Faust (1994) discuss this extensively about how this design introduced potentially severe measurement errors. This impacted the study of signed networks as well. Until recently, most of the analyses done with signed networks had, roughly, the same number of positive and negative ties. This had major consequences. The measures of imbalance and the use of the criterion function, C_f, defined above made sense if the number of negative and positive ties are the same. But what if they are not? If better network data collection methods are used or if signed graphs are studied in other empirical domains, the number of positive and negative ties could differ, sometimes greatly. For example, in addition

to producing the plot shown in Figure 3, Doreian & Mrvar (2015) found the number of positive ties far exceeded the number of negative ties. Hitherto, the line index of imbalance and the proportion of imbalanced triples had corresponded closely. Yet, in this network, they found that the line index of balance and the proportion of balanced triples did not correspond well: They pointed to different stories about changes in imbalance. Instead, the *number* of imbalanced triples corresponded well to the line index of imbalance. The presence of so many positive triples dramatically affected the classical measure of imbalance in signed networks, a problem that had not been recognized previously.

With the fixed choice design for collecting social network data having been abandoned largely – but not completely, alas – and with researchers considering other types of signed networks where the number of positive and negative ties differ considerably, serious problems have arisen. Particularly acute is the choice of α in the criterion function for signed blockmodeling. The idea of using C_f with α as a parameter was prompted by wanting to introduce some flexibility. It implied some consideration of which value of α to use for a given network. But this was not explored in the literature. Yet, with the above considerations, using $\alpha = 0.5$, while seemingly reasonable, now seems quite arbitrary. Other values for the parameter merit attention. Examining the use of different values of α has started. Initial experiments with changing α have shown that the blockmodel partitions of signed networks can differ with the values of α used as well as the line index measure of imbalance. The selection of αin a principled fashion is a major unsolved problem. Until it is solved, a question mark will hang over using signed blockmodeling. Using the proportion of imbalanced triples as a measure can be questioned also. Solving the 'alpha variation problem' will not be easy.

In terms of data, further issues arise. One is that we are able now to collect valued signed data. This poses no problems for the Doreian-Mrvar algorithms as implemented in Pajek. Yet the ways in which the values of the signed ties are operationalized in different instruments may matter. This merits further attention also. A more troublesome issue is the presumption that the signs of ties are either positive or negative (or null) when measured. Cartwright & Harary (1970) raised the issue of considering ambivalent relations that are real but neither positive nor negative. Such ambivalence could be quite stable. Or it could point to the temporal instability of ties that could be positive at one point in time but negative at another. Incorporating ambivalence into the study of signed relations is another problem for both substantive and methodological reasons. Certainly, the current signed blockmodeling approach of Doreian and Mrvar cannot handle ambivalence. Ignoring this is no longer an option – yet another difficult problem to solve. This problem is addressed, but only partially, in Mendonça *et al.* (2015) in their study of voting in the European Parliament with abstentions from voting. As they note, the meaning of 'abstaining' is not clear. Their response took the form of examining alternative ways of coding this tie. In human relations between individuals this approach may not work so well as ambivalence implies both a positive and a negative tie between actors – or a qualitatively different type of tie.

All the problems outlined in the previous four sub-sections appear to imply a serious consideration of the ways in which data are collected. On the one hand, implicit in the foregoing is a demand for obtaining more and higher quality data. Yet, collecting data can be expensive and there are considerations regarding respondent fatigue in single instruments and, even more, in longitudinal studies. A very good resource for designing research to obtain useful data is Robins (2015). There, studies must be designed carefully from the outset to their conclusion. Given the difficulties of developing high quality data, it is not surprising that obtaining and using data electronically is so appealing. While such data can be obtained within well designed studies, convenience often dominates as the main criterion for obtaining data. When this happens most, if not all, of the considerations regarding substance described above appear irrelevant. While we are clearly in an era of Big Data, I have major reservations about

the adequacy of these data when substance is ignored. That data exist does not imply they are worth analyzing simply because they exist. Of course, this observation can be extended to small data also!

3.5Algorithms

Assuming clean data can be collected for signed human relations, there are issues concerning the algorithms we use that must be addressed. The examples of the UNGA voting and Supreme Court voting described above suggest that, too often, our algorithms are just sledgehammers unfit for the subtleties of data in specific networks. The partitions established by Doreian & Mrvar (2015) were established after a painstaking examination of data once it was clear that well-known algorithms for partitioning such data failed. Of course, a painstaking look at the data may be necessary despite the risk of over-fitting the models to the data. Yet, it would be a major advance if efficient algorithms could be established for discerning the overall structure of such a network without the risk of over-fitting. Some research on this is under way but it is far from complete. One effort involves an ongoing collaboration between Traag, Doreian and Mrvar for a volume on partitioning networks to be published by Wiley. Clearly, Esmailian *et al.* (2014) have engaged on this issue as well with their consideration of relaxed structural balance. There is a well-known book review by Bonacich (2004) with the provocative title "The invasion of the physicists". While he was careful to note that the 'social' social network community has something to learn from this invasion, it struck a chord. Clearly, members of the two fields do need to engage far more with each other.

The shift from partitioning signed networks using structural balance to using relaxed structural balance was compelling for substantive and empirical reasons. The move to partitioning two-mode signed networks was logical and very useful. Yet, there was a cost. For the criterion function defined above for classical structural balance, Doreian *et al.* (2005, p.305) established a theorem stating there would be a unique lowest value of the criterion function for structural balance that would occur either for one number, k, of clusters in the partition or a set of adjacent values of k. Loosely, the criterion had a U-shape with a unique minimum value for the criterion function.

For relaxed structural balance, Doreian & Mrvar (2009, p.5) proved that, if the number of clusters is denoted by k, the criterion function declines monotonically as k increases. For signed two-mode networks, with k_1 and k_2 the number of clusters of the rows and columns respectively, Mrvar & Doreian (2009, p.204) proved the criterion function declines monotonically with k_1 for each value of k_2 and monotonically with k_2 for each value of k_1. The last two theorems imply that establishing partitions with a unique lowest value of the criterion function is not possible when the number of clusters is much lower than the number of vertices in the network. This ushers in judgment calls as to which partition, if any, is the best one to be selected – unless ways can be established to determine an optimum solution. Usually, the value of the criterion function declines faster for low values of k, k_1 and k_2 than for higher values of these parameters. In addition, to using judgment, there is an issue of efficiency in choosing the grain of a partition (choice of k). The more general problem is that if we think of the plot of the criterion function against k (for one-mode data), or against k_1 and k_2 (for two-mode data), in a three-dimensional plot, the shape of the plot is completely unknown. Worse, it may differ dramatically across the many different signed networks we could study. Therefore, the partitioning of the UNGA data mentioned above was so difficult. Characterizing such plots in general may be intractable as a formal problem. The work of Mendonça *et al.* (2015) suggests one way forward.

IVSUMMARY AND SOME PROVISIONAL CONCLUSIONS

This reflection has provided a partial summary of one consistent approach to the study of signed relations that has been around for about 70 years. While it outlined some successes,

clear failures were considered also. A closer look at these failures revealed a host of further major problems, both substantive and methodological. To build on the successes and, more consequentially, address the failures, these problems must be solved. Unless they are solved, the so-called balance theoretic approach will stall for there are looming potential dead ends. This adds urgency to pursuing the problems outlined here. No doubt, this must be done, especially as there are serious implications for developing adequate theories, designing both empirical and simulation studies of signed networks, collecting good data and employing well-designed algorithms for analyzing good data from different contexts.

At the risk of being viewed as a complete contrarian, I am not convinced of the value of focusing on the equilibrium states of signed networks. Given that signed networks move towards balance and away from balance, it seems highly unlikely that there will be long-term equilibria for them. Signed networks are highly dynamic. Equilibria appear to exist only in simulations studies.

As noted at the outset, there are multiple approaches to signed data. It may well be that ideas in some of these other approaches will be useful for solving some of the problems considered here. Clearly, Aref *et al.* (2016) provide such an example. It is possible, also, there are some ideas within the balance approach that will help solve some problems in these other approaches. See Mendonça *et al.* (2015) and Esmailian *et al.* (2014). More generally, sharing ideas across multiple approaches and between different disciplines seems the best way forward. It will be a massive quest but, I hope, not one that turns out to be quixotic. An academic friend, who changed sub-fields regularly, once remarked to me about research "When the easy problems have been solved, it is time to move on." I disagree: Once the easy problems have been solved, research life gets far more interesting. There are some juicy and interesting problems to chew upon when studying signed networks. It will certainly be an exciting and fun ride for those pursuing it.

References

Abell P., Ludwig M. (2009). Structural balance, a dynamic perspective.*Journal of Mathematical Sociology*, 33(2):129–155.doi: 10.1080/00222500902718239.

Aref S., Mason A.J., Wilson M.C. (2016). An exact method for computing the frustration index in signed networks using binary programming. arXiv:1611.09030v1 [cs:SI].

Batagelj V., Ferligoj A., Doreian P. (1992a). Direct and Indirect Methods for Structural Equivalence.*Social Networks*14(1-2):63-90.doi: 10.1016/0378-8733(92)90014-X.

Batagelj V., Doreian P., Ferligoj A. (1992b). An Optimizational Approach to Regular Equivalence.*Social Networks*14(1-2):121-135.doi: 10.1016/0378-8733(92)90016-Z.

Batagelj V., Mrvar A. (1998). Pajek - A Program for Large Network Analysis.*Connections* 21(2):47-57.

Bonacich P. (2004). The invasion of the physicists.*Social Networks* 26(3):285–288.doi: 10.1016/j.socnet.2004.06.002.

Breiger, R. L., Boorman, S. A., and Arabie, P. (1975). An algorithm for clustering relational data with applications to social network analysis and a comparison with multidimensional scaling. Journal of *Mathematical Psychology* 12(3):328-383.doi: 10.1016/0022-2496(75)90028-0.

Breiger R. L., Ennis J. G. (1979). Personae and social roles: The network structure of personality types in small groups.*Social Psychology Quarterly 42*(3):262-270.

Brusco M. J., Doreian P., Mrvar A., Steinley D. (2011). Two algorithms for relaxed structural balance partitioning: Linking theory, methods and data to understand social network phenomena.*Sociological Methods and Research* 40(1):57-87. doi: 10.1177/0049124110384947.

Burt R. S. (1976). "Positions in social networks", *Social Forces 55*:93-122.

Cartwright D., Harary F. (1956). Structural balance: A generalization of Heider's theory.*Psychological Review* 63:277-292.doi: 10.1037/h0046049.

Cartwright D., Harary F. (1970). Ambivalence and indifference in generalizations of structural balance.*Behavioral Science 15*(6):497–513.doi: 10.1002/bs.3830150604.

Davis J. A. (1967). Clustering and structural balance in graphs.*Human Relations* 20(2):181-187. doi: 10.1177/001872676702000207.

Doreian P. (2002).Event sequences as generators of social network evolution.*Social Networks*24(2):93-119.doi:10.1016/S0378-8733(01)00051-X.

Doreian P., Batagelj V., Ferligoj A. (2005). Generalized Blockmodeling. New York: Cambridge University Press.

Doreian P., Krackhardt D. (2001). Pre-transitive balance mechanisms for signed networks.*Journal of Mathematical Sociology* 25(1):43-67.doi: 10.1080/0022250X.2001.9990244.

Doreian P., Lloyd P., Mrvar A. (2013).Partitioning large signed two-mode networks: Problems and prospects.*Social Networks* 35(2):178-203.doi: 10.1016/j.socnet.2012.01.002.

Doreian P., Mrvar A. (1996). A partitioning approach to structural balance.*Social Networks*18(2):149-168.doi: 10.1016/0378-8733(95)00259-6.

Doreian P., Mrvar A. (2009). Partitioning signed social networks.*Social Networks* 31(1):1-11.doi: 10.1016/j.socnet.2008.08.001.

Doreian P., Mrvar A. (2014). Testing Two Theories for Generating Signed Networks Using Real Data.*Metodološki Zvezki: Advances in Methodology and Statistics* 11(1), 31-63

Doreian P., Mrvar A. (2015). Structural balance and signed international relations.*Journal of Social Structure* 16 Article 2.

Esmailian P., Abtahi S. E., Jalili M. (2014). Mesoscopic analysis of online social networks: The role of negative ties. Physical Review 90(4):042817.doi: 10.1103/PhysRevE.90.042817.

Harary F., Norman R. Z., Cartwright D. (1965). Structural Models. New York: John Wiley and Sons.

Heider F. (1946). Attitudes and cognitive organization.*Journal of Psychology* 21(1):107-112.doi: 10.1080/00223980.1946.9917275.

Hummon N. P., Doreian P. (2003). Some dynamics of social balance processes: Bringing Heider back into balance theory.*Social Networks* 25(1):17-49.doi: 10.1016/S0378-8733(02)00019-9.

Hummon N. P., Fararo T. J. (1995). Assessing hierarchy and balance in dynamic networks models.*Journal of Mathematical Sociology* 20(2-3):145-159.doi: 10.1080/0022250X.1995.9990159.

Kalish Y., Robins G.L. (2006). Psychological predispositions and network structure: Therelationship between individual predispositions, structural holes and network closure.*Social Networks* 28(1):56-84.doi: 10.1016/j.socnet.2005.04.004.

Lorrain F., While H. C. (1971). Structural equivalence of individuals in social networks.*Journal of Mathematical Sociology* 1(1):49-80.doi: 10.1080/0022250X.1971.9989788.

Mendonça I., Figueredo R., Labatut V., Michelon P. (2015). Relevance of negative links in graph partitioning: A case study using votes from the European Parliament.In*2nd European Network Intelligence Conference*, Karlskrona, Sweden, p.122-129.doi: 10.1109/ENIC.2015.25.

Mrvar A., Doreian P. (2009). Partitioning Signed Two-Mode Networks.*Journal of Mathematical Sociology* 33(3):196-221.doi: 10.1080/00222500902946210.

Newcomb T. M. (1961). The Acquaintance Process. New York: Holt, Rinehart, & Winston.

Nordlie P. (1958). A Longitudinal Study of Interpersonal Attraction in a Natural Setting. Unpublished Ph. D. Dissertation, University of Michigan.

Pevehouse J., Nordstrom T., Warnke K. (2004). The Correlates of War 2: International governmental organizations data, Version 2.0.*Conflict Management and Peace Science* 21(2):101-109.doi: 10.1080/07388940490463933.

Prota L., Doreian P. (2016). Finding roles in sparse economic hierarchies: going beyond regular equivalence.*Social Networks* 45:1-17.doi: 10.1016/j.socnet.2015.10.005.

Robins G. (2015). Doing Social Network Research: Network-based Research Design for Social Scientists. Los Angeles: Sage.

Robins G., Kashima Y. (2008). Social psychology and social networks: Individuals and social systems.*Asian Journal of Social Psychology* 11(1):1-12.doi: 10.1111/j.1467-839X.2007.00240.x.

adamsj., Faust K., Lovasi G. S. (Eds) (2012). Capturing Context: Integrating Spatial and Social Network Analyses.*Social Networks*34(1):1-158, special issue.

Taylor H. F. (1970). Balance in Small Groups. New York: Van Nostrand Reinhold.

Traag V.A., Bruggeman J. (2009).Community detection in networks with positive and negative links.*Physical Review E* 80(3):036115. doi:10.1103/PhysRevE.80.036115.

Vinogradova G., Galam S. (2014).Global alliances effect in coalition forming.*The European Physical Journal B* 87:266.doi: 10.1140/epjb/e2014-50264-4.

Wasserman S., Faust K. (1994). Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, USA.

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

Analyse des Préférences et Tournois Pondérés

Alain GUÉNOCHE

Institut de Mathématiques de Marseille (AMU - CNRS)

*Correspondance : alain.guenoche@univ-amu.fr

DOI : 10.18713/JIMIS-170117-2-2
Soumis le 31 mai 2016 - Publié le 17 janvier 2017
Volume : **2** - Année : **2017**
Titre du numéro : **Graphes & systèmes sociaux**
Éditeurs : *Rosa Figueiredo & Vincent Labatut*

Résumé

Dans de nombreuses études expérimentales, on dispose de n éléments ordonnés suivant plusieurs classements (votes, notes ou critères). Nous traitons et comparons deux problèmes : (i) Etablir un classement unique (ordre total) des n items et (ii) sélectionner les k meilleurs éléments parmi n. Il s'agit, dans les deux cas, de minimiser le nombre de préférences qui vont à l'encontre de ces choix.

Mots-Clés

Préférences ; Tournoi ; Ordres médians ; Sélection de candidats

I INTRODUCTION

Les données de préférences sont induites par des ordres :

- des ordres totaux, c'est à dire des classements sur un ensemble d'individus ou d'items, tous classés, qui n'admettent pas d'ex-aequo ;
- des préordres totaux, qui admettent les ex-aequo. Ce sont donc des ordres totaux sur des classes d'équivalence, constituées des items à égalité du point de vue de l'ordre.

Ces ordres peuvent venir de votes (opinions formulées par des classements sur des individus ou des jugements d'appréciation sur des produits). Ils peuvent aussi provenir de notes attribuées aux items par des critères quantitatifs ou des épreuves comparatives. On retrouve ainsi les problèmes liés aux procédures de dépouillement de scrutins, dont l'origine remonte à Condorcet (Caritat, 1785), aux classements hédonistes basés sur des appréciations d'experts, à la sélection de candidats pour des concours de recrutement ou au classement des élèves selon plusieurs matières.

Nous adoptons les formulations suivantes : soit X un ensemble d'items ($|X| = n$) et Π un ensemble de m ordres ou préordres totaux, appelé *profil* [1]

$$\Pi = \{O_1, O_2, \ldots, O_m\}. \quad (1)$$

Les préférences sont les paires ordonnées d'items. On notera $x \prec_O y$ si x est avant y dans O. Un ordre est équivalent à l'ensemble de ses préférences et la distance *naturelle* entre les ordres O_i et O_j est la distance de Kendall δ :

$$\delta(O_i, O_j) = |\{(x, y) \in X^2 \text{ tels que } (x \prec_{O_i} y \wedge y \prec_{O_j} x) \vee (y \prec_{O_j} x \wedge x \prec_{O_i} y)\}|. \quad (2)$$

Nous abordons deux problèmes :

— (i) Etablir un ordre *consensus* le plus en accord possible avec le profil, c'est-à-dire un ordre total π sur X ;
— (ii) Sélectionner une partie de X, S de cardinal k fixé. La classe complémentaire est notée $R = X - S$, R pour rejetés.

Le principe est de minimiser le nombre de préférences des ordres du profil Π qui (i) contredisent celles de π ou (ii) celles qui vont de R vers S.

Un ordre qui minimise la somme des préférences du profil en contradiction avec cet ordre est appelé un ordre *médian*. C'est tout à fait justifié dans le cas d'un profil d'ordres totaux car, pour la distance de Kendall δ et si $T(x, y)$ est le nombre d'ordres du profil qui placent x avant y. Cet ordre consensus π est une *médiane* du profil car $\sum_{i=1,\ldots,m} \delta(O_i, \pi)$ est minimum dans l'ensemble des ordres totaux sur X (Barthélemy et Monjardet, 1981). Mais si le profil contient des préordres, ou si les pondérations des arcs ne sont pas les nombres d'ordres, un ordre à écart minimum du tournoi n'est pas assuré d'être une médiane. Néanmoins, on continuera d'employer, de manière abusive, le terme *ordre médian* pour tout *ordre à écart minimum d'un tournoi*. S'il n'est pas unique, on cherchera à énumérer tous les ordres optimaux. Pour le second problème, on construit un groupe de taille k fixée, sans chercher à ordonner les items à l'intérieur des classes.

Du point de vue méthodologique, nous allons voir que ces deux problèmes sont différents. Le premier est classique dans le cadre de la *Théorie du choix social* (vote) ou de l'*Agrégation des préférences* selon des comparaisons par paires (Barthélemy et Monjardet, 1981). Le problème de sélection est souvent traité via un ordre total, en retenant les k éléments qui sont les mieux classés. Nous verrons que, du point de vue de la minimisation du nombre de préférences contradictoires, c'est une erreur !

Ces problèmes peuvent être posés en termes de programmation mathématique en variables entières, puisqu'il s'agit d'affecter des positions ou des rangs à n items qui introduisent des pénalités si leurs positions relatives ne respectent pas les préférences majoritaires. On voit ainsi que ces problèmes sont des cas particuliers du fameux problème d'affectation quadratique (QAP) pour lequel de nombreuses heuristiques d'approximation et des méthodes Branch & Bound ont été développées (Finke *et al.*, 1987). L'obtention d'une borne inférieure du coût d'une affectation partielle est souvent réalisée par des méthodes de programmation mathématique (Roupin, 2004), ce qui n'est pas le cas dans notre approche.

Les méthodes développées ci-dessous reposent sur des comparaisons par paires des éléments de X. De chaque ordre ou préordre du profil, on retient que x est meilleur que y, ou l'inverse

1. en référence aux profils d'opinions.

ou encore qu'ils sont à égalité. Notons $T(x, y)$ le poids des préférences de x en faveur de y. Le plus souvent, $T(x, y)$ est le nombre d'ordres dans Π qui préfèrent x à y [2]. Si $T(x, y) > T(y, x)$ la préférence de x en faveur de y est dite *majoritaire*, ce qui correspond à la terminologie des votes.

Au profil Π on associe un *tournoi* (graphe complet orienté de sommets X) dont les arcs correspondent aux préférences majoritaires. Entre les sommets x et y on met un arc (x, y) de x vers y si et seulement si $T(x, y) > T(y, x)$. Cet arc a pour poids la différence $w(x, y) = T(x, y) - T(y, x)$ et, dans la table W des poids, on pose $w(y, x) = 0$. Si $T(x, y) = T(y, x)$, on peut mettre l'arc dans n'importe quel sens, puisque $w(x, y) = w(y, x) = 0$.

Tout ordre total sur les items $O = (x_1, x_2, \ldots, x_n)$ distingue deux types d'arcs ; les *arcs directs* qui sont orientés selon l'ordre O et les *arcs-retour* qui sont orientés dans le sens opposé. Soit $W(O)$ la somme des poids de ces arcs-retour qui mesure l'*écart* de l'ordre O au tournoi. Cet ordre est d'autant plus compatible avec le tournoi que $W(O)$ est faible [3]. Mais que faut-il compter dans W ?

— Dans le premier problème, on considère tous les arcs-retour et on cherche donc un ordre total sur X qui minimise

$$W_o(O) = \sum_{i<j} w(x_j, x_i). \tag{3}$$

— Dans le second cas, on cherche seulement deux classes, une bipartition orientée, avec d'un côté les sélectionnés S et de l'autre les rejetés R et on ne compte que les arcs-retour qui vont de R vers S.

$$W_p(S|R) = \sum_{x \in S, y \in R} w(y, x) \tag{4}$$

Pour construire des ordres médians ou une classe d'items sélectionnés, on a recourt à des méthodes d'énumération. Pour améliorer leur efficacité il est nécessaire de disposer de bornes supérieures aux optimima que l'on cherche à atteindre. Elles sont calculées à l'aide d'ordres dont l'écart au tournoi est optimisé de façon approchée. Dans la Section II, on part de deux heuristiques classiques pour introduire une optimisation locale et une méthode basée sur une décomposition du tournoi. Elles sont originales et permettent d'améliorer toute solution approchée.

Dans la Section III nous reprenons la méthode *Branch & Bound* développée par Guénoche (1977); Barthélemy *et al.* (1989); Charon *et al.* (1996), et nous l'adaptons au problème de l'énumération de tous les ordres médians. L'étude d'un petit tournoi, correspondant au choix de conférenciers invités lors d'un colloque et qui admet plusieurs ordres médians, nous amène à étudier la question du choix d'un de ces ordres. La Section IV est dédiée au problème de sélection de k éléments selon des ordres, que l'on résout aussi par énumération.

II DES HEURISTIQUES POUR UN "BON" ORDRE

La construction d'un ordre optimal à partir d'un tournoi résulte d'une méthode *Branch & Bound* qui nécessite la connaissance d'une borne supérieure de l'écart d'un ordre au tournoi. Ces

2. On peut aussi tenir compte de l'écart entre x et y, la différence des rangs ou, dans le cas de mesures quantitatives, des différences entre les notes de x et de y. Dans ce cas, $T(x, y)$ est la somme des différences positives entre les valeurs de x et de y et une procédure de normalisation des notes est préconisée

3. Un tournoi qui admet un ordre O tel que $W_o(O) = 0$ est dit *transitif*.

bornes sont indispensables pour établir un ordre médian, car elles permettent de limiter la taille de l'arborescence de recherche. Nous ne reprenons que deux heuristiques classiques, la méthode gloutonne et une méthode de score avant de présenter nos améliorations. La méthode de Smith et Payne (1974) (retournement de l'arc qui supprime le plus de 3-circuits), étendue aux tournois pondérés (Barthélemy *et al.*, 1989) s'est avérée sans efficacité pour les problèmes pondérés, surtout quand leur taille n augmente.

Il existe d'autres approches qui relèvent de l'optimisation stochastique, à commencer par le Recuit Simulé (Metropolis), l'optimisation en voisinage variable (Hansen et Mladenović, 2001) ou la méthode de bruitage (Charon et Hudry, 2001). Les paramètres à prendre en compte, les temps de calcul et les codes propres à leurs auteurs font que nous ne les considérons pas dans cette étude. Bien évidemment, toute solution heuristique peut servir de point de départ à toute procédure d'optimisation.

2.1 La méthode gloutonne

Elle procède du même principe que la procédure *Branch & Bound*, sauf qu'elle ne développe pas d'arborescence des sections commençantes et choisit à chaque itération l'item qui promet un ordre de moindre coût. On calcule donc les sommes colonnes de la table des poids du tournoi $Sum(y) = \sum_{x \in X} w(x, y)$, contribution de y à l'écart d'un ordre commençant par x.

i=0 ; Tant qu'il existe un élément non placé

- soit x non placé tel que $Sum(x)$ est minimum
- i++ ; $o_i = x$; x est placé ;
- pour tout y non placé
  - $Sum(y) = Sum(y) - w(x, y)$;

Cette heuristique est clairement en $O(n^2)$.

2.2 Les méthodes de score

Ce sont des méthodes basées sur des fonctions de score. Chaque ordre attribue des points à chaque item et le score d'un item est la somme éventuellement pondérée, des points acquis. Suivant la signification des points, l'ordre consensus est l'ordre croissant ou décroissant des scores. L'exemple type est la procédure de Borda (1784), rival de Condorcet, pour désigner les vainqueurs d'une élection. Son score est le rang du candidat (1 point pour le premier, 2 pour le second, etc) et l'ordre final est l'ordre croissant de la somme des rangs. Dans le cas de préordres, il faut attribuer à chaque item le rang moyen des ex-aequo.

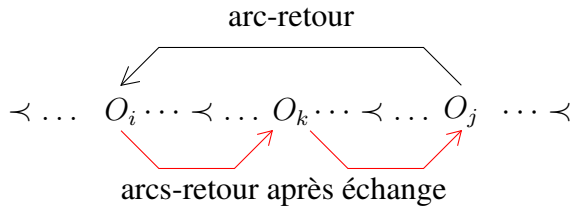
Un autre exemple classique est le championnat des pilotes de Formule 1 ; chaque course attribue des points (non proportionnels) aux dix premiers, et 0 pour les autres. Le classement est l'ordre décroissant de la somme des points acquits par les pilotes.

2.3 Une optimisation locale

Etant donné un ordre $O = (o_1, o_2, \ldots, o_n)$ sur X, la procédure d'optimisation locale cherche par échange d'items un ordre d'écart au tournoi plus petit. Pour tout élément o_j, soit o_i le dernier élément placé avant lui tel que $w(o_j, o_i) > 0$; on a donc $i < j$, et (o_j, o_i) est le plus petit arc-retour partant de o_j. Le cas où o_i et o_j sont consécutifs correspond à une procédure très classique : en transposant ces deux éléments, on diminue l'écart au tournoi, puisqu'on supprime un arc-retour. Sinon, on a intérêt à échanger o_j et o_i si et seulement si les items placés entre o_i et o_j ne créent pas d'arcs-retour de poids plus grand. C'est ce que l'on vérifie en sommant les valeurs des arcs-retour après échange.

Soit $O = (o_1, o_2, \ldots, o_n)$ un ordre total.
Tant qu'il y a échange

— Pour j=2 à n
  — soit (o_j, o_i) le plus petit arc retour d'origine o_j
  — $Som = \sum_{i>k>j} w(o_k, o_j) + w(o_i, o_k)$
  — si $w(o_j, o_i) > Som$ échanger o_i et o_j



Ainsi l'intervalle $(o_i, o_{i+1}, \ldots, o_{j-1}, o_j)$ devient $(o_j, o_{i+1}, \ldots, o_{j-1}, o_i)$. Cette procédure est appliquée itérativement tant qu'il existe des transpositions de ce type à réaliser. Pour tout item, il suffit de remonter vers le dernier item précédent dominé et d'effectuer des sommations dans l'intervalle ainsi défini. Chaque itération est en $O(n^2)$ dans le cas le pire (celui de l'ordre induit par un tournoi transitif pour lequel on cherche en vain un intervalle à retourner).

Dans la suite de ce texte, nous appelons *BestH* la meilleure des deux heuristiques après application de la procédure d'optimisation locale.

2.4 Décomposition d'un tournoi

Devant des problèmes de grande taille ($n \gg 30$), les heuristiques classiques sont peu efficaces. D'où l'idée de décomposer le tournoi en sous-tournois, c'est à dire partitionner les items de façon que chaque classe contiennent des éléments voisins dans un "bon" ordre. Il suffit alors de calculer un ordre consensus pour chaque classe et de les concaténer pour obtenir un ordre total, certes approché, mais plus proche du tournoi que les heuristiques globales. Nous montrons, à l'aide de simulations sur des tournois aléatoires, que cette statégie de décomposition est efficace.

2.4.1 Une décomposition régulière

L'ordre donné par l'une des heuristiques précédentes induit une décomposition régulière. Etant donné un nombre de classes q, il suffit de construire les classes comme des intervalles de cet ordre. Les n/q premiers vont dans la première classe, les n/q suivants dans la seconde, etc. On obtient ainsi une partition en classes équilibrées, notée P_R.

2.4.2 Une partition des items basée sur une distance

En considérant la table W des poids des arcs du tournoi, on peut associer à chaque item x une bipartition : Soit x_+, les éléments de X qui devraient être placés avant x parce qu'ils le dominent et x_-, qui devraient être placés après x parce qu'il les domine.

$$x_+ = \{z \in X | w(z, x) > 0\} \text{ et } x_- = \{z \in X | w(x, z) > 0\}. \tag{5}$$

A l'aide de ces bipartitions, on construit un indice de dissimilarité sur X,

$$D(x, y) = \Delta(x_+, y_+) + \Delta(x_-, y_-), \tag{6}$$

dans lequel Δ est la distance de la différence symétrique : $\Delta(A,B) = |\{x \in \{A \backslash B \bigcup B \backslash A\}\}|$. On notera que D n'est pas une distance, car il se peut que $D(x,y) = 0$ si $w(x,y) = w(y,x) = 0$).

Proposition Si T est un tournoi transitif,

— deux éléments consécutifs dans l'ordre optimal ont une distance égale à 2 ;

— $D(x,y)$ est proportionnelle à la différence des rangs de x et y dans l'ordre médian.

Preuve

Soit $x \prec y$ deux éléments consécutifs dans l'ordre correspondant à un tournoi transitif. Les classes x_+ et y_+ (resp. x_- et y_-) ne diffèrent que d'un seul élément, x (resp. y) et donc $D(x,y) = 2$. De même, si x et y sont séparés par k items dans l'ordre médian, $D(x,y) = 2(k+1)$. Ainsi les valeurs sont croissantes quand on s'écarte de la diagonale, et D est une distance (car il n'y a pas d'indifférence dans les préférences).

Ainsi, les faibles valeurs de D correspondent à des éléments voisins dans un ordre médian et les classes obtenues doivent contenir des éléments consécutifs. Le choix du nombre de classes q sera conduit par une procédure de simulation. Pour construire, à partir de D une partition en q classes, notée P_M, on cherche à minimiser la somme des distances intra-classe par une procédure itérative à la manière des k-means. On part de la partition atomique qui ne contient que des singletons. A chaque itération un item est transféré vers la classe pour laquelle sa moyenne de distance est minimum. La procédure s'arrête quand il n'y a plus d'item à déplacer.

Algorithme de décomposition

Soit $O = (o_1, o_2, \ldots, o_n)$ un ordre total sur X

1. Découper X en q classes ordonnées $\{X_1, \ldots, X_q\}$
   — soit en classes consécutives équilibrées suivant O
   — soit en classes calculées suivant la distance D
2. Calculer un ordre consensus O_i pour chaque classe X_i
3. Concaténer les ordres $O_1, \ldots, O_q$ suivant le rang moyen des éléments de X_i dans O
4. Appliquer la procédure d'optimisation locale

On obtient ainsi un *ordre composé* noté $Comp_R$ (resp. $Comp_M$), suivant la méthode de partitionnement choisie, et son écart au tournoi est calculé.

2.4.3 Complexité, efficacité

Le calcul de la table des $O(n^2)$ valeurs de distances est en $O(n^3)$, puisque pour chaque paire d'items on compare les positions relatives de n éléments. L'algorithme de décomposition régulière est linéaire. La partition P_M est obtenue par un algorithme itératif, dont on ignore le nombre d'itérations (comme pour les k-means), mais qui est bien connu pour son efficacité. Après, on applique l'heuristique gloutonne à chaque classe et la procédure d'optimisation à l'ordre composé.

Notons que la méthode de l'*Ordre composé* est rapide. Il faut compter 1"20 pour P_R, 19"30 pour P_M sur un tournoi à 1000 items sur un MacPro portable.

n	m	q	BestH	$Comp_R$	$Comp_M$
50	10	2	153	152	**151**
50	20	3	259	251	**249**
50	30	4	330	319	**318**
100	10	3	755	731	**719**
100	20	3	1236	1203	**1187**
100	30	3	1569	1534	**1522**
200	30	4	6825	6678	**6614**
200	50	5	9047	8820	**8784**
200	100	6	13129	**12762**	12782
500	100	5	86600	85336	**85107**
500	100	10	86600	**84670**	85635
500	100	15	86600	**84636**	86078

TABLE 1 – Valeurs moyennes des écarts obtenus par les heuristiques sur des tournois générés par des permutations aléatoires.

2.4.4 Simulations et résultats

Sur des profils aléatoires

On génère des profils d'ordres totaux en tirant au hasard des permutations d'ordre n (Nijenhuis et Wilf, 1975). Chaque profil permet de calculer le tournoi avec les poids des arcs (matrice W). Après application de l'optimisation locale, on dispose de deux ordres totaux. La meilleure des solutions, notée BestH, est celle qui minimise l'écart au tournoi.

On se fixe alors le nombre de classes q. Pour le partitionnement des items, on calcule D avec les bipartitions $P_x = (x_+|x_-)$. On a :

— la partition régulière (P_R) en suivant l'ordre donné par BestH qui aboutit à l'ordre $Comp_R$;
— une partition (P_M) qui cherche à minimiser les moyennes des distances intra classes et qui aboutit à l'ordre $Comp_M$;

Pour BestH et chacun de ces ordres, on calcule leur écart au tournoi.

Les tests portent sur 100 profils de mêmes paramètres ; à nombre d'items fixé n, on peut faire varier le nombre d'ordres totaux dans le profil m et le nombre de classes de la décomposition q. Chaque ligne de la Table 1 donne l'écart moyen aux tournois ainsi générés.

Il est clair que les ordres composés sont meilleurs que la meilleure des heuristiques classiques. Soulignons que ce sont des écarts moyens et que, pour un profil ou pour un tournoi particulier, les deux décompositions peuvent être testées. Mais on ne sait pas quelle méthode est la plus efficace, ni combien de classes assurent un écart minimum. La partition P_M semble donner les meilleurs résultats, mais peut être que le nombre de classes est trop petit, comme le suggère le nombre croissant de classes pour $n = 500$.

Sur des tournois à écart borné

En tirant au hasard des ordres totaux indépendants, on génère des tournois qui sont loin d'être transitifs et donc les ordres obtenus sont d'écart élevé. Et il est impossible de savoir s'ils sont proches de l'optimum. D'un certain point de vue, il est inutile de chercher un ordre consensus à ces profils, puisqu'aucune tendance commune ne peut émerger. Nous traîtons des tournois

n	m	t	q	$BestH$	$Comp_R$	$Comp_M$	$OrdNat$
100	10	20	5	81.7	73.1	**66.3**	111.9
100	20	20	5	22.6	21.1	**19.5**	23.9
100	30	20	5	5.3	**5.1**	5.2	5.9
300	30	100	10	2065.0	1721.0	**1434.0**	1555.0
300	50	100	10	800.0	677.0	**558.0**	544.0
300	100	100	10	56.1	50.6	**49.1**	45.9

TABLE 2 – Valeurs moyennes des écarts obtenus par les heuristiques sur des tournois à écart borné.

générés à partir d'un ordre unique en transposant un certain nombre d'items. Soit t le paramètre qui définit le nombre de transpositions.

On génère m ordres totaux à partir de l'ordre naturel en échangeant, dans chacun, t paires d'items aléatoirement choisis. Le tournoi pondéré est alors calculé selon la procédure majoritaitre classique. Si t est faible par rapport à n, la tendance générale est alors l'ordre naturel, qui a toutes les chances d'être médian.

L'essai porte sur deux familles d'ordres ; des ordres à 100 (resp. 300) items, sur lesquels on effectue 20 (resp.100) transpositions avant de calculer le tournoi pondéré. Les décompositions sont en 5 (resp. 10) classes.

Plus le nombre d'ordres est grand, plus le tournoi est transitif et l'ordre naturel proche de l'optimal. La seconde méthode de décomposition s'avère très efficace quand on s'en éloigne.

2.5 Conclusion

La stratégie de décomposition est toujours gagnante et donc il est préférable d'assembler des ordres obtenus sur les sous-tournois en un ordre unique, plutôt que de traiter le tournoi d'un seul bloc. De plus, pour des tournois "presque transitifs", les ordres obtenus par décomposition restent proches de l'ordre optimal.

Donc, pour un tournoi particulier, je commencerai par chercher le nombre optimal de classes à l'aide du découpage régulier qui est très rapide, puisqu'il n'y a même pas de distance à calculer. Et autour de cette valeur, je testerai les algorithmes de partitionnement. Un dernier essai sur un tournoi à 1000 items a donné par une décomposition régulière en 15 classes le plus faible écart, et ce très nettement.

III ENUMÉRATION DES ORDRES OPTIMAUX

Construire un ordre total π dont l'écart au tournoi $W_o(\pi)$ est minimum (le problème de Kemeny (1959)), revient à rendre le tournoi transitif par retournement d'arcs, ceux dont la somme des poids est minimum. Ce problème, *minimum feedback arc set*, a une longue histoire algorithmique, qui débute avec Slater (1961); Remage et Thomson (1966); Bermond (1972) pour les tournois non pondérés, puis tout un groupe centré autour du Centre d'Analyse et de Mathématiques Sociales (CNRS, EHESS) avec Monjardet (1973); Guénoche (1977); Barthélemy *et al.* (1989); Charon *et al.* (1996); Hudry (2012) pour les tournois pondérés.

Depuis Hudry (1989), le problème a été prouvé NP-difficile. Le calcul s'effectue par une méthode de *Branch & Bound* (Guénoche, 1977) dont nous rappelons tout d'abord le principe. On développe une arborescence dont les sommets sont des *sections commençantes*. Ce sont des débuts d'ordres

totaux qui peuvent se prolonger en un ordre total optimal. C'est à dire que le poids de cette section est au plus égal à la borne supérieure B_o calculée par l'une des heuristiques précédentes.

Une feuille de l'arborescence courante est donc une chaîne $(o_1, o_2, \ldots, o_k)$ sur une partie $X_s \subset X$ et notons $Y = X - X_s$ l'ensemble des éléments non classés. Le poids W_o^k de cette section est égal à la somme des poids des arcs-retour internes à cette section plus la somme des poids des arcs dont l'origine y est hors cette section et l'extrémité dans cette section.

$$W_o^k(o_1, o_2, \ldots, o_k) = \sum_{1 \leq i < j \leq k} w(o_j, o_i) + \sum_{y \in Y} \sum_{j=1,\ldots,k} w(y, o_j). \tag{7}$$

Cette valeur est inférieure ou égale au poids de tout ordre total ayant cette section commençante, puisqu'elle ne tient pas compte de l'ordre non encore déterminé sur Y.

En prolongeant l'arborescence, à chaque itération, par la feuille de poids minimum, on aboutit à un ordre total sur X dont l'écart est minimum. Sa valeur est déterminée par le premier ordre total obtenu. Si le tournoi contient beaucoup de circuits, cette méthode peut devenir inapplicable et ce dès que n atteint quelques dizaines d'items (Barthélemy *et al.*, 1989). Pour deux raisons ; l'une est la taille de l'arborescence qui peut dépasser le million de noeuds [4] et l'autre le temps de calcul, à cause de la recherche de la feuille de l'arborescence de poids minimum que l'on prolonge d'abord pour être sûr d'aboutir à un premier ordre total optimum. Cette difficulté a été contournée par Woirgard (1997) en adaptant une structure de données plus élaborée.

De façon plus efficace, pour obtenir tous les ordres médians, au lieu de développer l'arborescence à partir d'une feuille de moindre poids (stratégie du meilleur d'abord), on la développe en largeur d'abord, ce qui évite d'avoir à chercher à chaque pas la feuille de poids minimum.

Algorithme d'énumération des ordres médians

Soit B_0 une borne supérieure de l'écart d'un ordre total au tournoi et A une arborescence initiale de racine $\emptyset$ à laquelle sont attachés les éléments de X tels que $Sum(x) = \sum_{y \in X} w(y, x) \leq B_0$.

Pour toute feuille S_k de A

1. $S_k = (o_1, o_2, \ldots, o_k)$ d'écart W_o^k ;
2. Pour tout y non placé dans S_k
   - $W_o^{k+1} = W_o^k + \sum_{z \in X - S_k} w(z, y)$
   - Si $W_o^{k+1} \leq B_0$ prologer A avec la section $(o_1, o_2, \ldots, o_k, y)$, attachée à S_k

La taille de l'arborescence dépend fortement de la qualité de B_o. De fait, on enumère tous les ordres totaux dont l'écart est inférieur à B_o, borne que l'on met à jour au fil de l'énumération, s'il y a lieu. On peut même fixer une borne inférieure à B_o pour laquelle l'arborescence n'aboutit à aucun ordre si elle est inférieure à l'écart minimum, et ce très rapidement.

4. Pour chaque noeud, selon une structure de données élaborée par Guénoche (1977), il suffit de mémoriser 3 entiers.

1. $(x_1 \prec x_2 \prec x_3 \prec x_4 \prec x_5, x_6, x_7)$
2. $(x_5 \prec x_1 \prec x_2 \prec x_4 \prec x_6 \prec x_3, x_7)$
3. $(x_7 \prec x_6 \prec x_2 \prec x_3 \prec x_1, x_4, x_5)$
4. $(x_5 \prec x_3 \prec x_4 \prec x_1, x_2, x_6, x_7)$
5. $(x_4 \prec x_7 \prec x_5 \prec x_6 \prec x_2 \prec x_1, x_3)$
6. $(x_6 \prec x_1 \prec x_7 \prec x_2 \prec x_3, x_4, x_5)$

TABLE 3 – Préordres totaux des 6 membres du comité de programme.

W	x_1	x_2	x_3	x_4	x_5	x_6	x_7
$x1$	-	1	1	1	0	0	1
$x2$	0	-	4	2	0	0	0
$x3$	0	0	-	1	0	0	0
$x4$	0	0	0	-	0	2	2
$x5$	1	0	1	0	-	1	0
$x6$	1	1	2	0	0	-	0
$x7$	0	1	1	0	1	0	-

TABLE 4 – Tournoi majoritaire résultant des préordres de la Table 3.

3.1 Le tournoi Jobim 2016

Nous prenons comme exemple tout au long de ce paragraphe le choix de conférenciers invités pour le colloque national de bioinformatique, Jobim. Chaque membre du comité de programme désigne les conférenciers qu'il voudrait faire venir et il les range dans un ordre préférentiel. Les membres du comité ne se sont pas trop dispersés et ont nommé 10 candidats ; nous avons retenu ceux qui étaient dans au moins la moitié des ordres. Il est resté $n = 7$ candidats (les 3 autres étant éliminés). Dans le profil, nous avons repris les ordres donnés. Pour chacun, on a enlevé les éliminés et ajouté, en fin de classement et tous ex-aequo, les candidats non retenus pour obtenir un préordre total.

Le tournoi majoritaire pondéré résultant de ces préférences est donné dans la Table 4.

On constate que tous les candidats sont dominés au moins une fois. Il n'y a donc pas de *vainqueur de Condorcet*. Pour déterminer un ordre total ou une bipartition de poids minimum on applique l'heuristique gloutonne, la taille du tournoi ne nécessitant pas une décomposition. Ici, le sommet le moins dominé est x_5 ; un ordre commençant par x_5 n'aura qu'un seul arc-retour de poids 1 (x_7, x_5). Ce sommet étant "placé", on retient x_1 qui n'entraîne qu'un autre arc-retour de poids 1 (x_6, x_1). Et ainsi de suite. On aboutit à l'ordre glouton :

$$O_g = (x_5 \prec x_1 \prec x_2 \prec x_4 \prec x_6 \prec x_7 \prec x_3), \tag{8}$$

et $W_o(O_g) = 5$ avec $\{(x_7, x_5), (x_6, x_1), (x_6, x_2), (x_7, x_2), (x_3, x_4)\}$, comme arcs-retour, tous de poids 1, donc $B_o = 5$. Un ordre optimal est au plus à écart 5 du tournoi.

La procédure de Borda appliquée aux préordres initiaux (en donnant un rang moyen aux candidats à égalité) conduit à deux paires d'ex-aequo $\{x_1, x_5\}$ et $\{x_4, x_7\}$. En respectant les orientations des arcs du tournoi, on aboutit à :

$$O_b = (x_2 \prec x_5 \prec x_1 \prec x_6 \prec x_4 \prec x_7 \prec x_3) \tag{9}$$

qui présente 6 arcs-retour et $W_o(O_b) = 7$, un écart supérieur à O_g.

1. $O_1 = (x_1 \prec x_2 \prec x_4 \prec x_7 \prec x_5 \prec x_6 \prec x_3)$
2. $O_2 = (x_1 \prec x_4 \prec x_7 \prec x_5 \prec x_6 \prec x_2 \prec x_3)$
3. $O_3 = (x_4 \prec x_5 \prec x_6 \prec x_1 \prec x_7 \prec x_2 \prec x_3)$
4. $O_4 = (x_4 \prec x_7 \prec x_5 \prec x_6 \prec x_1 \prec x_2 \prec x_3)$
5. $O_5 = (x_5 \prec x_1 \prec x_2 \prec x_4 \prec x_6 \prec x_7 \prec x_3)$
6. $O_6 = (x_5 \prec x_1 \prec x_2 \prec x_4 \prec x_7 \prec x_6 \prec x_3)$
7. $O_7 = (x_5 \prec x_1 \prec x_4 \prec x_6 \prec x_7 \prec x_2 \prec x_3)$
8. $O_8 = (x_5 \prec x_1 \prec x_4 \prec x_7 \prec x_6 \prec x_2 \prec x_3)$
9. $O_9 = (x_5 \prec x_4 \prec x_6 \prec x_1 \prec x_7 \prec x_2 \prec x_3)$
10. $O_{10} = (x_5 \prec x_6 \prec x_1 \prec x_2 \prec x_3 \prec x_4 \prec x_7)$
11. $O_{11} = (x_5 \prec x_6 \prec x_1 \prec x_2 \prec x_4 \prec x_7 \prec x_3)$
12. $O_{12} = (x_5 \prec x_6 \prec x_1 \prec x_7 \prec x_2 \prec x_3 \prec x_4)$
13. $O_{13} = (x_7 \prec x_5 \prec x_6 \prec x_1 \prec x_2 \prec x_3 \prec x_4)$

TABLE 5 – Ordres médians à écart 5 du tournoi de la Table 4.

La procédure de recherche d'un ordre médian aboutit à un autre ordre O_m à écart 5, ce qui confirme l'optimalité de B_o.

$$O_m = (x_4 \prec x_7 \prec x_5 \prec x_6 \prec x_1 \prec x_2 \prec x_3). \tag{10}$$

On lance donc la procédure d'énumération qui donne les 13 ordres médians suivant :

3.2 Simulations

Est-ce que ces difficultés sont fréquentes ou exceptionnelles ? Pour en avoir le coeur net, nous avons tiré au hasard des profils d'ordres totaux. Pour obtenir des chiffres crédibles pour les invités d'un colloque, nous avons fixé à $n = 20$ le nombre de candidats et à $m = 31$ le nombre d'experts, un nombre impair pour éviter les égalités et donc les arcs de poids nul.

Le plus simple est de partir d'ordres aléatoires, de calculer le tournoi majoritaire, puis d'énumérer tous les ordres médian et de compter le nombre moyen de solutions en répétant 100 fois ce tirage. Mais en tirant des ordres aléatoires, on simule une situation bien improbable ; qu'il n'y ait pas de consensus véritable entre les experts non plus que de renommée relative entre les candidats. Pour graduer ce degré d'entente entre les experts, nous sommes partis de l'ordre naturel et simulons les votes en effectuant t échanges de deux items tirés au hasard, pour tous les ordres. Ainsi, si $t = 0$ tout le jury est d'accord, et quand t croît on s'approche du désordre aléatoire qui correspond à $t = 30$.

Pour ces quatre jeux d'essais, $t = 10, 15, 20, 30$, nous tirons donc 100 profils et comptons la valeur moyenne d'écart donnée par l'heuristique gloutonne (O_g), l'ordre de Borda (O_b), un ordre médian (O_m), le nombre moyen d'ordres optimaux ($NbOpt$), le nombre maximum ($NbMax$), la taille moyenne de l'arborescence ($Tree$), ainsi que son maximum ($Tree_{Max}$).

La différence entre l'écart donné par l'heuristque et l'écart optimal parait faible, mais il y a beaucoup d'ordres entre ces deux valeurs, ce qui souligne la supériorité (en moyenne) de O_g sur O_b. Le nombre moyen de solutions optimales reste limité, mais le nombre maximum observé est important et justifie le paragraphe ci-dessous pour décider d'un classement unique des candidats. Enfin la taille de l'arbre est importante, mais les calculs sont très rapides, quelques secondes pour les 100 problèmes les plus difficiles.

t	O_g	O_b	O_m	$NbOpt$	$NbMax$	$Tree$	$Tree_{Max}$
10	9.9	14.8	9.2	2.1	18	187	6695
15	23.2	27.2	20.5	3.3	28	2116	53544
20	35.0	38.7	30.2	4.0	40	12331	964554
30	42.5	47.0	36.7	4.1	40	47766	1391501

TABLE 6 – Ecart moyen des heuristiques, d'un ordre médian, les nombres moyens et maximum d'ordres optimaux et les nombres moyens et maximum de nœuds de l'arbre de recherche.

3.3 Choisir parmi les ordres médians

En présence de plusieurs ordres médians on aimerait n'en garder qu'un pour établir un ordre consensus unique. La tâche n'est pas facile. Sur les 13 ordres optimaux du tournoi Jobim, il y a 4 premiers, x_1, x_4, x_5 et x_7, dont deux d'entre eux peuvent être derniers. Tous les candidats, sauf x_3, sont dans les deux premiers rangs. Voici quelques solutions :

3.3.1 Le plus central des ordres médians

De tous ces ordres, quel est le plus central ? C'est celui dont la somme des distances aux autres ordres médians est minimum. Il suffit de calculer la distance de Kendall (nombre de paires d'items en désaccord entre deux ordres) et d'en faire les sommes ; c'est O_7 le plus central.

3.3.2 Le plus souvent dans les k premiers

Si l'on avait un seul invité, on prendrait x_5 parce qu'il est le plus souvent le premier dans l'ensemble des ordres médians. Et en seconde position vient celui qui, en dehors de x_5 est le plus souvent premier ou second ; c'est x_1. Et ainsi de suite ; on place au rang k un candidat non placé qui est le plus souvent (plus que les autres) à un rang inférieur ou égal à k. Cette procédure fonctionne comme un scrutin uninominal à un tour qui est répété k fois. Ici, on aboutit encore au même ordre O_7.

Ce n'est pas équivalent à la procédure de Borda appliquée aux ordres médians, même si ici elle donne le même résultat, car on ne tient aucun compte des mauvais classements. Par contre on n'aboutit pas nécessairement à un ordre médian, c'est pourquoi je préconise l'autre solution au choix d'un ordre unique.

IV A LA RECHERCHE D'UNE SÉLECTION

Dans ce paragraphe, nous cherchons à sélectionner les k meilleurs éléments parmi n en minimisant le nombre de préférences qui vont à l'encontre de ce choix. Précisons qu'ici le paramètre k est fixé. Les arcs-retour entre les sélectionnés S ou entre les rejetés R ne comptent pas, puisqu'ils ne contredisent en rien la sélection finale. On cherche donc deux classes et on ne compte que les arcs-retour qui vont de la classe R vers S.

$$W_p(S|R) = \sum_{x \in S, y \in R} w(y, x) \tag{11}$$

Entre deux classes complémentaires de sommets dans un graphe orienté, on parle de *cocircuit*. On cherche donc un cocircuit dont les parties sont de cardinal fixé et dont le poids des arcs-retour est minimum.

Du point de vue méthodologique, ce problème est différent de celui de l'ordre consensus. Souvent, le problème de sélection est traité via un classement de tous les items, en retenant les

W	x_1	x_2	x_3	x_4	x_5	x_6
$x1$	-	5	0	1	1	1
$x2$	0	-	5	1	1	1
$x3$	5	0	-	1	1	1
$x4$	0	0	0	-	1	1
$x5$	0	0	0	0	-	1
$x6$	0	0	0	0	0	-

TABLE 7 – Tournoi pour illustrer la non monotonie de W_p^k quand k augmente.

k items les mieux classés. Nous verrons que, pour minimiser W_p, c'est une erreur ! Comme précédemment, un ordre total sur X dont l'écart au tournoi soit faible est nécessaire car il fournit une borne supérieure au critère W_p. Cet ordre permet de fixer le poids minimum d'un cocircuit d'arcs-retour sur une classe S de cardinal k, notée B_p^k.

Pour illustrer ce paragraphe, nous reprenons le tournoi des conférenciers invités au colloque Jobim, avec $k = 2$. Examinons tout d'abord les ordres donnés par les heuristiques classiques :

Selon l'ordre glouton $O_g = (x_5 \prec x_1 \prec x_2 \prec x_4 \prec x_6 \prec x_7 \prec x_3)$, une bipartition avec deux invités, x_1 et x_5 donnent $W_p^2(x_1, x_5|x_2, x_3, x_4, x_6, x_7) = 2$ (avec les arcs-retour $\{(x_6, x_1), (x_7, x_5)\}$, donc $B_p^2 = 2$.

L'ordre de Borda $O_b = (x_2 \prec x_5 \prec x_1 \prec x_6 \prec x_4 \prec x_7 \prec x_3)$ suggère une partition à deux invités, avec $W_p^2(x_2, x_5|x_1, x_3, x_4, x_6, x_7) = 4$, le double du précédent.

4.1 Une sélection optimale

Ce problème semble encore plus difficile [5] que celui de l'ordre médian, car la construction de cette bipartition n'est pas cumulative. Bien sûr, on est tenté de prendre comme "premier" invité le candidat le moins dominé. Mais il se peut qu'un candidat plus dominé puisse être retenu par la suite avec celui qui le domine fortement. La cause en est la non monotonie de la fonction de poids W_p quand le nombre de sélectionnés augmente, alors que W_o l'était quand on prolongeait les sections commençantes.

Prenons l'exemple du petit tournoi pondéré de la Table 7 :

S'il n'y a qu'un invité, c'est x_4 de poids 3, s'il y en a deux, c'est l'une des trois paires $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$ de poids 5. Mais si l'on en prend trois, c'est $\{x_1, x_2, x_3\}$ car cette partie domine tous les autres et $W_p^3(x_1, x_2, x_3|x_4, x_5, x_6) = 0$.

Donc, si l'on veut un seul item, il suffit de comparer les sommes colonnes de la table du tournoi. Pour 2 ou 3 éléments, il faut effectuer quelques boucles imbriquées. Mais si l'on veut une partie à $k > 3$ éléments, je ne vois pas de méthode optimale autre que l'énumération des parties à k éléments parmi N.

5. Non pas au sens de la théorie de la complexité, mais du calcul pratique

4.2 Des bornes efficaces

Heureusement, cette énumération ne porte pas nécessairement sur toutes les parties de X à k éléments. D'abord un ordre total construit à l'aide d'une heuristique fournit une borne supérieure B_p^k au poids du cocycle retour. Ensuite, pour chaque item x, on peut établir deux bornes, $B_S(x)$ (resp. $B_R(x)$) correspondant aux poids minimum des arcs-retour si $x \in S$ (resp. $x \in R$). Si x est sélectionné ($x \in S$), il peut être extrémité d'éventuels arcs-retour. Il y en a au plus $N-k$ dont le poids est supérieur ou égal à la somme des $N-k$ plus petites valeurs des arcs d'extrémité x, c'est à dire les $N-k$ plus petites valeurs de la colonne x de la table W :

$$B_S(x) = \sum_{i=1,\dots N-k} Min\{w(y_i, x)\}_{y_i \neq x}. \tag{12}$$

De même, si x est rejeté ($x \in R$), il peut être à l'origine de k arcs-retour, soit au minimum la somme des k plus petites valeurs de la ligne de x.

$$B_R(x) = \sum_{i=1,\dots k} Min\{w(x, y_i)\}_{y_i \neq x}. \tag{13}$$

Ces deux bornes se calculent facilement par tri des lignes et des colonnes de W. Elles permettent souvent d'affecter x à l'une ou l'autre classe si B_S ou B_R dépassent la borne supérieure B_p^k connue pour une bipartition avec k sélectionnés. Comme nous le verrons dans les simulations ci-dessous, c'est ce qui advient pour plus de la moitié des candidats. On peut donc procéder à une énumération sur un nombre restreint d'items parmi lesquels on ne cherche plus qu'un nombre petit de candidats pour compléter ceux qui sont sélectionnés d'office.

Pour une énumération efficace des parties de cardinal fixé d'un ensemble, Nijenhuis et Wilf (1975) proposent un algorithme qui, pour passer d'une partie à la suivante, supprime et ajoute un seul élément. Ainsi, le poids de la nouvelle partie peut être calculé à partir du poids de l'ancienne en $O(n)$. Mais la complexité de cette énumération reste liée à $\binom{n}{k}$.

Pour le tournoi Jobim, s'il n'y a qu'un seul invité, x_5 est le seul à reçevoir un arc de poids minimum. Si l'on veut deux invités, x_1 et x_5 n'en reçoivent que deux, et c'est la seule paire qui réalise ce score. Mais s'il y a trois invités, on trouve deux triplets de score 3, $\{x_1, x_5, x_6\}$ et $\{x_1, x_5, x_7\}$; le second n'est même pas un début d'ordre médian !

4.3 Simulations again

Ici aussi, nous avons voulu (i) comparer les heuristiques, (ii) tester la difficulté du calcul et (iii) savoir si les bornes permettent d'affecter directement des candidats. Pour les mêmes tournois que précédemment, avec 31 juges et 20 candidats, nous cherchons à sélectionner 5 invités. Nous mesurons le poids moyen des cocircuits-retour obtenus par l'heuristique gloutonne, par l'ordre de Borda, puis l'énumération fournit la valeur optimale. Nous indiquons de plus le nombre moyen de candidats dont l'affectation est fixée par les bornes ($NbFix$), le nombre moyen de solutions optimales et le maximum obtenu sur 100 tirages.

L'heuristique gloutonne est proche de l'optimum et fait nettement mieux que la procédure de Bordat. On notera qu'a contrario des ordres, il y a peu de bipartitions optimales, le plus souvent une seule, même pour des classements aléatoires. De façon imprévue, l'énumération porte sur moins de la moitié des candidats, parfois aucun car ils sont tous affectés d'office.

t	O_g	O_b	O_{opt}	$NbFix$	$NbSol$	$NbMax$
10	1.8	4.1	1.7	18	1.2	3
15	5.6	7.1	5.3	15	1.2	3
20	9.3	10.8	8.6	12	1.2	3
30	11.4	12.8	10.7	11	1.2	3

TABLE 8 – Valeurs moyennes des poids des cocircuits-retour, du nombre moyen d'item fixés et des nombres moyen et maximum de classes optimales.

V CONCLUSIONS

Dans ce texte, nous avons essentiellement réalisé trois tâches nouvelles :

— calculé de meilleures bornes supérieures de l'écart d'un ordre total à un tournoi grâce à l'optimisation locale et à la décomposition du tournoi ;
— séparé les deux problèmes de classement (ordre total) et de sélection d'un nombre fixé de candidats ;
— proposé une méthode optimale de sélection dans le cas où n est petit, à l'aide de bornes efficaces.

D'une manière générale, nous préconisons la démarche suivante [6] :

— S'entendre sur les candidats, à l'issue d'un premier choix libre ou d'une discussion entre les membres du comité de programme, et sur le nombre d'invités ;
— Faire voter le comité qui exprime un ensemble d'ordres totaux, sur les candidats, si possible en évitant les ex-aequo ;
— Construire le tournoi pondéré par les préférences majoritaires ;

Pour le problème de construction d'un ordre unique ;

— Enumérer toutes les ordres médians ;
— Choisir le plus central de ces ordres.

Pour le problème de sélection de $k > 3$ candidats ;

— Calculer les bornes supérieure d'appartenance à l'une ou l'autre classe ;
— Enumérer les bipartitions des candidats dont l'affectation n'est pas fixée.

La principale vertu de cette approche est qu'elle tient au mieux compte des classements initiaux, puisqu'elle contredit un nombre minimum de préférences individuelles et qu'on ne peut faire moins. Même si les algorithmes sont de complexité élevée ils s'appliquent à des données de cette taille, courantes dans bien des situations évoquées dans l'introduction. Pour la sélection, si on ne veut qu'un seul ou deux ou trois sélectionnés les solutions optimales sont toujours calculables.

La procédure d'énumération ne demande pas de place mémoire et peut s'envisager pour des problèmes plus importants. Nous obtenons un groupe de 10 sélectionnés parmi 30 candidats, avec 100 votants assez divergents, puisque les ordres ont été générés avec 50 transpositions aléatoires, en moins d'une minute de calcul. Il y a de quoi choisir les meilleurs crus de l'année, selon un pannel d'experts, ou les "entrants" d'une section du CNRS, autres domaines d'application de cette méthode de sélection.

6. Les programmes sont disponibles sur simple demande à l'auteur.

Références

Barthélemy J. P., A. G., Hudry O. (1989). Median linear orders : Heuristics and branch and bound algorithm. *European Journal of Operational Research 42*(3), 555–579. doi:10.1016/0377-2217(89)90442-6.

Barthélemy J. P., Monjardet B. (1981). The median procedure in cluster analysis and social choice theory. *Mathematical Social Sciences 1*(3), 235–267. doi:10.1016/0165-4896(81)90041-X.

Bermond J. C. (1972). Ordre à distance minimale d'un tournoi et graphes partiels sans circuits maximaux. *Mathématiques et Sciences humaines 37*, 5–25.

Borda J. C. (1784). Mémoire sur les élections au scrutin. In *Histoire de l'Académie Royale des Sciences pour 1781–1784*, pp. 657–664. Imprimerie Royale.

Caritat M. J. A. N. Marquis de Condorcet. (1785). *Essai sur l'application de l'analyse de la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale.

Charon I., Guénoche A., Hudry ., Woirgard F. (1996). A bonsaï branch and bound method applied to voting theory. In *Ordinal and Symbolic Data Analysis*, pp. 309–318.

Charon I., Hudry O. (2001). The noising methods : a generalization of some metaheuristics. *European Journal of Operational Research 135*(1), 86–101. doi:10.1016/S0377-2217(00)00305-2.

Finke G., Burkard R. E., Rendl F. (1987). Quadratic assignment problems. *Annals of Discrete Mathematics 31*, 61–82.

Guénoche A. (1977). Un algorithme pour pallier l'effet Condorcet. *R.A.I.R.O. Recherche Opérationnelle 11*, 77–83.

Hansen P., Mladenović N. (2001). Variable neigborhood search : Principles and applications. *European Journal of Operational Research 130*(3), 449–467. doi:10.1016/S0377-2217(00)00100-4.

Hudry O. (1989). *Recherche d'ordres médians : complexité, algorithmique et problèmes combinatoires*. Thèse de doctorat, École nationale supérieure des télécommunications, Paris.

Hudry O. (2012). On the computation of median linear orders, of median complete preorders and of median weak orders. *Mathématical Social Sciences 64*(1), 2–10. doi:10.1016/j.mathsocsci.2011.06.004.

Kemeny J. G. (1959). Mathematics without numbers. *Daedalus 88*(4), 577–591.

Monjardet B. (1973). Tournoi et ordre médian pour une opinion. *Mathématiques et Sciences humaines 43*, 55–73.

Nijenhuis A., Wilf H. S. (1975). *Combinatorial algorithms*. Academic Press.

Remage R., Thomson W. A. (1966). Maximum likelihood paired comparison rankings. *Biometrica 53*(1/2), 143–149. doi:10.2307/2334060.

Roupin F. (2004). From linear to semidefinite programming : an algorithm to obtain semidefinite relaxations for bivalent quadratic problems. *Journal of Combinatorial Optimization 8*, 469–493. doi:10.1007/s10878-004-4838-6.

Slater P. (1961). Inconsistencies in a schedule of paired comparisons. *Biometrika 48*(3/4), 303–312. doi:10.2307/2332752.

Smith A. F. M., Payne C. D. (1974). An algorithm for determining Slater's i and all nearest adjoining orders. *British Journal of Mathematical and Statistical Psychology 27*(1), 49–52. doi:10.1111/j.2044-8317.1974.tb00526.x.

Woirgard F. (1997). *Recherche et dénombrement des ordres médians des tournois*. Thèse de doctorat, École nationale supérieure des télécommunications, Paris.

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

Brazilian Congress structural balance analysis

Mario Levorato[*1] **and Yuri Frota**[1]

[1]Department of Computer Science, Fluminense Federal University, Brazil

*Corresponding author: mlevorato@ic.uff.br

DOI: 10.18713/JIMIS-280217-2-3
Submitted: September 5 2016 - Published: February 28 2017
Volume: **2** - Year: **2017**
Issue: **Graphs & Social Systems**
Editors: *Rosa Figueiredo & Vincent Labatut*

Abstract

In this work, we study the behavior of Brazilian politicians and political parties with the help of clustering algorithms for signed social networks. For this purpose, we extract and analyze a collection of signed networks representing voting sessions of the lower house of Brazilian National Congress. We process all available voting data for the period between 2011 and 2016, by considering voting similarities between members of the Congress to define weighted signed links. The solutions obtained by solving Correlation Clustering (CC) problems are the basis for investigating deputies voting networks as well as questions about loyalty, leadership, coalitions, political crisis and polarization.

Keywords

Social Network; Signed Graph; Structural Balance; Correlation Clustering; Metaheuristic; Politics

I INTRODUCTION

Structural balance theory is based on the notion of cognitive consistency between friendship and hostility. For example, an enemy of a friend is probably my enemy as well, while a friend of a friend is probably my friend or can become one (Heider, 1946). In simple terms, the interaction of individuals follows the tendency to create stable (albeit not certainly conflict-free) social groups. This can be specially interesting to study similarity and correlation networks, like those originated from common voting patterns, or alliances and disputes among parties or nations (Traag and Bruggeman, 2009; Macon *et al.*, 2012; Doreian and Mrvar, 2015).

One appropriate criterion to measure the degree of balance in signed social networks is by solving the Correlation Clustering (CC) problem (Bansal *et al.*, 2002; Demaine *et al.*, 2006), which consists of partitioning a set of elements into clusters by analyzing the level of similarity between them. It aims to maximize the affinity inside each cluster (i.e. positive relationships)

while, at the same time, minimizing the similarities between elements of different clusters (i.e. maximizing negative relationships).

The CC problem, which has been proved to be NP-hard (Bansal *et al.*, 2002), can be applied in several areas, such as efficient document classification (Bansal *et al.*, 2002), natural language processing (Elsner and Schudy, 2009), image segmentation (Kim *et al.*, 2014) and, of course, signed social network analysis (Doreian and Mrvar, 1996; Brusco *et al.*, 2011; Figueiredo and Moura, 2013; Levorato *et al.*, 2015). With this objective, the level of balance in a social group can be used by social network researchers to study how (and if) a group evolves to a possible balanced state.

A relaxed version of the CC problem called Symmetric Relaxed Correlation Clustering (SRCC) problem (Brusco *et al.*, 2011; Figueiredo and Moura, 2013) can also be used to evaluate balance in signed social networks. This variant, although computationally harder to solve, allows the identification of special types of social relationship, such as polarization, mediation and differential popularity (Doreian and Mrvar, 2009), originally viewed as violations of structural balance.

We implemented an algorithm known as $ILS{-}CC$ (Levorato *et al.*, 2015), which can efficiently solve the aforementioned problems, providing useful information for social network analysis. Using the House of Cunha website (Andrade, 2016) and the work of Mendonça *et al.* (2015) as inspiration, we provide a novel analysis of Brazilian politics inside the Chamber of Deputies (CD). In Brazil, the Chamber of Deputies (*Câmara dos Deputados*) is the lower house of the National Congress, comprised of 513 federal deputies (from 25 political parties), elected by a proportional representation of votes to serve a four-year term. Based on the CD voting records, we generate several instances of signed social networks, according to certain grouping criteria. The clustering results obtained when invoking the $ILS - CC$ procedure over these instances are the starting point of our study.

The analysis presented in this work can be applied to any network originated from voting patterns, where alliances and interest groups have strong influence.

This paper is organized as follows. Section II presents a literature review regarding Correlation Clustering problems and signed social network analysis. Section III describes the method applied to extract signed networks from the Chamber of Deputies voting data. Section IV presents an analysis of structural balance on the Chamber of Deputies voting networks, based on the solutions obtained by using our methodology. Finally, we show our conclusions in Section V.

II RELATED WORKS

Heider (1946) was the first to state Structural Balance (SB) theory in order to define sentiment relations among people belonging to the same social group (such as like/dislike, love/hate and cooperation/competitivity). Signed graphs were later applied by Cartwright and Harary (1956), formalizing SB theory which affirmed that a stabilized social group could be divided into two mutually hostile subgroups (or clusters), each having internal solidarity. Davis (1967) then proposed the more general notion of "weak balance" or clusterable signed graph, when a balanced social group can be divided into two or more mutually antagonistic subgroups, each having internal solidarity.

When solving a clustering problem, one wants to find the most balanced partition[1] of a signed

[1] A partition is here defined as the division of the set of vertices V into non-overlapping and non-empty subsets.

graph. Using structural balance as a measure, the clustering problem is equivalent to solving the optimization problem called Correlation Clustering (CC). To our knowledge, this problem was first addressed by Doreian and Mrvar (1996) (although not under this name), who provided a heuristic solution method for analyzing structural balance on real-world social networks. Their method was implemented in software Pajek (De Nooy *et al.*, 2011). Having a document clustering problem in mind, Bansal *et al.* (2002) formalized the unweighted version of the CC problem and also discussed its NP-completeness proof. Later, Demaine *et al.* (2006) addressed the weighted version of the problem. Integer linear programming (ILP) can be used to solve the CC problem optimally (Figueiredo and Moura, 2013), but only if the number of elements is small. Since it consists of a NP-hard minimization problem, the only available solutions for larger instances are either heuristic or approximate. The solution of the CC problem and of some of its variants has already been applied in several areas, such as portfolio analysis in risk management (Harary *et al.*, 2002; Huffner *et al.*, 2009), biological systems (DasGupta *et al.*, 2007; Huffner *et al.*, 2009), grouping of genes (Bhattacharya and De, 2008), efficient document classification (Bansal *et al.*, 2002), image segmentation (Kim *et al.*, 2014) and community structure (Traag and Bruggeman, 2009).

In Yang *et al.* (2007), the CC problem is known as *community mining* and an agent-based heuristic called FEC is proposed to obtain its solution. Genetic algorithms have also been applied to document clustering, using the CC problem as objective function (Zhang *et al.*, 2008). Lately, Drummond *et al.* (2013) presented a Greedy Randomized Adaptive Search Procedure (GRASP) (Feo and Resende, 1995) implementation that provides an efficient solution to the CC problem in networks of up to 8000 vertices. Then, based on this method, Levorato *et al.* (2015) introduced sequential and parallel ILS (Iterated Local Search) (Lourenço *et al.*, 2003) procedures for the CC problem (known as $ILS - CC$), which outperformed other solution methods from the literature on three huge real-world signed social networks. Similarly to Mendonça *et al.* (2015), in this work, we will use the $ILS - CC$ algorithm to evaluate the imbalance of voting networks.

Apart from the CC problem, alternative measures to structural balance and the associated clustering problems have also been discussed in the literature. In Doreian and Mrvar (2009), the definition of a k-balanced signed graph was informally extended in order to include relevant processes (polarization, mediation, differential popularity and subgroup internal hostility) that were originally viewed as violations of structural balance. For example, the existence of a group of individuals who share only positive relationships with everyone in the network counts as imbalance in the CC Problem. Nonetheless, the individuals in this group could be identified as mediators (i.e. their relations probably won't change over time) and, as pointed in Doreian and Mrvar (2009), their relations should not be considered as a contribution to the imbalance of the network.

Using this new definition, structural balance was generalized to a version labeled as *relaxed structural balance* (Doreian and Mrvar, 2009). Similarly to the CC problem, measuring the relaxed structural balance can be accomplished through the solution to the Relaxed Correlation Clustering (RCC) problem. It is originally defined on asymmetric relations between clusters (Figueiredo and Moura, 2013); however, a redefinition of relaxed imbalance of a partition P that takes into account only symmetric relationships is also available. This gives rise to a new graph clustering problem, the Symmetric Relaxed Correlation Clustering (SRCC) Problem (Figueiredo and Moura, 2013), which will be used in this work. The SRCC problem allows us to analyze mediation processes (positive and negative). That is not the case of the RCC

problem, where mediation and differential popularity cannot be pointed out.

It is worth noting that the SRCC problem is closely related with the CC problem but it is not a particular case nor is a generalization. Actually, each feasible solution (a graph partition) of the SRCC problem is also feasible in the CC problem but the problems have different cost functions, i.e., there are different ways of evaluating the imbalance of a partition. The SRCC problem is intuitively as difficult as the CC problem and is indeed a NP-hard problem (Figueiredo and Moura, 2013).

Two solution methods were initially presented in the literature for RCC problems: a greedy heuristic approach (Doreian and Mrvar, 2009) and a branch-and-bound procedure (Brusco *et al.*, 2011). Computational experiments with both procedures were reported over literature instances with up to 29 vertices and for random instances with up to 40 vertices (Doreian and Mrvar, 2009; Brusco *et al.*, 2011). We extended the ILS procedure to solve the SRCC problem, by applying additional data structures and a new objective function to evaluate the partition (Levorato *et al.*, 2017). As far as we know, the $ILS - CC$ algorithm is the only metaheuristic approach that has been used to solve RCC problems.

Previous works have employed signed graph clustering methods to analyze networks of international alliances and disputes (Traag and Bruggeman, 2009; Macon *et al.*, 2012; Doreian and Mrvar, 2015). In Levorato *et al.* (2015), by using the $ILS - CC$ algorithm, we presented a historical and geopolitical analysis of the results obtained from the voting on resolutions in the United Nations General Assembly (UNGA). Mendonça *et al.* (2015) have then applied a parallel version of the $ILS - CC$ algorithm to analyze a collection of signed networks representing voting sessions of the European Parliament. The obtained results were compared to a selection of community detection algorithms designed to process only positive links.

Several authors studied the voting behavior of politicians. As far the European Parliament (EP) is concerned, Hix (2002) evaluated different questions regarding voting behavior in the EP, including personal policy preferences, national party and European party disciplines. Afterwards, Hix and Noury (2009) compared the voting behavior of Members of the European Parliament (MEPs) in different periods, analyzing issues such as party cohesion and coalition formation.

In particular, regarding Brazil, Ames (1995) developed a model of legislative voting based on the operation of Brazil's political institutions. Mainwaring and Liñán (1997) studied party discipline in the Brazilian constitutional congress of 1987–88. Figueiredo and Limongi (2000) investigated how Brazilian presidents have succeeded by relying on the support of disciplined parties in order to get their agendas approved in the Congress. Calvão *et al.* (2015) performed an extensive analysis of data sets available for Brazilian proportional elections of legislators and city councilors throughout the period of 1970–2014, plus a comparative analysis of elections for legislative positions, in different states and years.

III NETWORK EXTRACTION

In this section, we explain the retrieval of raw voting data, and how we extracted signed networks from it.

3.1 Brazilian Chamber of Deputies

The Chamber of Deputies (CD) provides web services[2] which supply information about each of its members, including the vote cast by a specific deputy for each proposition evaluated at the CD. A deputy is described by its name, state (one of 27 Brazilian Federative Units) and political party.

For a given proposition, a deputy can express his vote in either of four ways (Câmara, 2016): *Sim* (For: the deputy wants the proposition to be accepted), *Não* (Against: s/he wants the proposition to be rejected), *Abstenção* (Abstain: s/he refuses to take part in the election and does not vote; equivalent to a white vote) and *Obstrução* (Filibuster: a form of obstruction, where debate over a proposition is extended, in order to delay or entirely prevent a vote on the proposal).

Besides the previous votes, a deputy may not vote at all, which leads to a fifth vote type: Ausência (Absent: the deputy was not present during the voting session).

The Chamber of Deputies' web services provide raw voting data, which describe the behavior of deputies apart from the others. Nonetheless, since a network is naturally relational (relationships between individuals are the product of their opinion about topics of interest), voting data has to be processed to generate the networks we wish to analyze.

3.2 Extraction algorithm

The extraction method here explained is based on the work of Mendonça *et al.* (2015). However, this procedure is being applied to Brazilian voting networks for the first time, which demanded an extension to the original algorithm, for filibuster treatment. It starts with a comparison between all pairs of deputies, analyzing the similarity of their voting choices. The obtained measures make up what is known as the agreement matrix M. Each element m_{uv} of this matrix indicates the average agreement between two deputies u and v, in other words, their level of accordance taking into consideration all propositions voted during a given time period.

While filtering the results is a relatively simple task, processing agreement scores may seriously alter the resulting network, depending on the methodology applied. Given a certain pair of deputies u and v and a proposition p_i, the proposition-wise agreement score $m_{uv}\ (p_i)$ is determined by comparing the votes of both deputies. It ranges from -1 if they fully disagree (one voted FOR and the other AGAINST), to +1 if they entirely agree (they share the same vote: FOR or AGAINST).

As previously stated, a voting record may contain, besides FOR and AGAINST, other values which should be equally taken into account. The first case refers to absence of one deputy or both of them (it is worth remembering that the analysis is based on pairs of deputies). The general approach is to leave out all propositions p_i that fall into this case (Porter *et al.*, 2005; Dal Maso *et al.*, 2014). Since certain deputies have low attendance rates, this might lead to distorted agreement or disagreement average scores, due to the small number of common voting sessions. To prevent this, we assume a neutral score of zero if at least one deputy is absent when voting a given proposition.

The abstention process is more complicated to understand. For example, if the political party supports a completely different view from the deputy, such pressure may be enough to lead

[2]Please visit http://www2.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo/webservices

him/her to take a step towards abstention, despite the fact that s/he is FOR or AGAINST the proposition under analysis. Similarly, abstention may simply represent the deputy's neutral position when a specific topic is proposed (i.e. the deputy does not care whether or not the subject is approved). Literature provides different views to deal with ABSTAIN-FOR, ABSTAIN-AGAINST and ABSTAIN-ABSTAIN situations (Macon *et al.*, 2012; Porter *et al.*, 2005; Dal Maso *et al.*, 2014). In this work, we make use of two different ways of calculating the scores. The first one (Table 1) treats abstention as half an agreement whenever it is paired with FOR, AGAINST or other abstention, yielding a value of +0.5. In the second one (Table 2), whenever two deputies abstain at the same time, this is viewed as a full agreement (+1 value). As opposed to that, if only one abstains, a zero score is assigned, since there is not sufficient information to assert they are in agreement or disagreement. So to make things more clear, absence was not included in the tables.

The last case is filibuster (or obstruction), a vote choice specific to the Brazilian Congress, which does not occur in the European Parliament and was, therefore, not studied by Mendonça *et al.* (2015). Such practice is used to create difficulties or hindrances in a systematic way to delay or impede the approval of a bill in parliament. It is normally used by minority groups which do not have the necessary number of representatives to effectively hold back a decision taken by the majority. Therefore, any vote marked as obstruction is here regarded as AGAINST.

	FOR	ABSTAIN	AGAINST
FOR	+1	+0.5	-1
ABSTAIN	+0.5	+0.5	+0.5
AGAINST	-1	+0.5	+1

Table 1: Vote weights representing abstention as half an agreement Mendonça *et al.* (2015).

	FOR	ABSTAIN	AGAINST
FOR	+1	0	-1
ABSTAIN	0	+1	0
AGAINST	-1	0	+1

Table 2: Vote weights representing abstention as absence of opinion Mendonça *et al.* (2015).

The proposition-wise agreement score is fully specified by choosing one of the previous processing strategies. By averaging this score over all considered propositions, the average agreement can be calculated (Mendonça *et al.*, 2015). In a formal way, consider two users u and v, as well as the propositions resulting from the filtering stage: p_1, ..., p_ℓ, for which both u and v voted. The average agreement m_{uv} between these two deputies is:

$$m_{uv} = \frac{1}{\ell} \sum_{i=1}^{\ell} m_{uv}(p_i) \tag{1}$$

Similarly to the work of Mendonça *et al.* (2015), we generated one signed graph for each year (from 2011 until June 2016), taking into account all the voting sessions in that year. Graph edges with weight smaller than 0.001 were removed from the graph. The set of vertices in each signed graph represents the list of deputies who voted at least one time in the corresponding year.

IV STRUCTURAL BALANCE ANALYSIS

In this section, based on the clustering results obtained with the ILS-CC algorithm on the graphs extracted according to Section 3.2, we investigate some aspects of Brazilian politics in the Chamber of Deputies, including leadership, polarization, coalitions, loyalty and government crisis.

As explained in the previous section, we followed two approaches when generating voting networks for each year in the period between January 2011 and June 2016. We will refer to each network as either v1 or v2, depending on the strategy while dealing with abstentions:

- **v1** : abstention is worth half an agreement (+0.5), whenever it is paired with any kind of vote (FOR, AGAINST or other abstention);
- **v2** : abstention is viewed as full agreement (+1 value) only if both deputies abstain. Otherwise, if only one abstains, a zero score is assigned.

In order to improve the readability of some charts, not all party labels have been displayed. For full information, all charts and tables used in this analysis are available on-line[3].

4.1 A brief introduction to Brazilian politics

From 1994 to 2002, Brazil was governed by president Fernando Henrique Cardoso, member of the PSDB (Brazilian Social Democracy Party). In 2002, PSDB was defeated in the presidential elections by PT (Brazilian Labor Party) and president Lula da Silva was elected for a four-year term, being reelected in 2006 for one more period of four years. Then, in 2010, president Dilma Rousseff (also a PT member and supported by president Lula da Silva) won the elections, becoming the next president and, like her predecessor, was also reelected in 2014 for an additional four-year term.

Since 2013, Brazil has been facing intense political and economical crisis, aggravated by successive scandals of corruption in the heart of the government (Connors, 2016; Robins-Early, 2016). In 2016, an impeachment process was started, on charges related to breaking budget laws, and president Dilma Rousseff was turned away from her post (Watts, 2016b; BBC, 2016). However, a more detailed research over international news articles reveals different views about the root causes of the political crisis and the impeachment itself (Alston, 2016; Bevins, 2016; Connors, 2016; Leahy, 2016; Rapoza, 2016; Shahshahani and Nation, 2016; Taub, 2016).

In order to help understanding the political groups and parties referenced in the analysis, we first provide a list of the three most voted candidates and their party alliances during the presidential elections held in 2010 (Table 3) and in 2014 (Table 4). In our analysis, we will refer to the first party alliance (candidate Dilma Rousseff, in both presidential elections) as the government coalition, while the second party alliance (candidates José Serra in 2010, and Aécio Neves in 2014) will be called opposition.

Candidate	**Coalition parties**	**#**
Dilma Rousseff	PCDOB, PDT, PMDB, PR, PRB, PSB, PSC, **PT**, PTC, PTN	10
José Serra	DEM, PMN, PPS, **PSDB**, PTB, PTDOB	6
Marina Silva	**PV***	1

Table 3: Major candidates in the 2010 presidential elections, ordered by the number of votes (column # contains the number of parties). Six more candidates (from six remaining parties) ran for presidency in 2010. (*) Like PV, their parties were not in a coalition.

[3]Please visit https://public.tableau.com/profile/mario.levorato

Volume: 2 - Year: 2017 , DOI: 10.18713/JIMIS-280217-2-3

Candidate	Coalition parties	#
Dilma Rousseff	PCDOB, PDT, PMDB, PP, PR, PRB, PROS, PSD, **PT**	9
Aécio Neves	DEM, PEN, PMN, **PSDB**, PTB, PTC, PTDOB, PTN, SD	9
Marina Silva	PHS, PPL, PPS, PRP, **PSB**, PSL	6

Table 4: Major candidates in the 2014 presidential elections, ordered by the number of votes (column # contains the number of parties). Eight more candidates (from eight remaining parties) ran for presidency in 2014. Their parties were not in a coalition.

Another useful piece of information is the list of parties according to their orientation (Table 5).

Orientation	Parties	#
Left	PCB, PCDOB, PCO, PSOL, PSTU, PT	6
Center-left	PDT, PMN, PPL, PPS, PROS, PSB, PSDB, REDE, SD	9
Center	DEM, PEN, PHS, PMB, PMDB, PRP, PSD, PSDC, PSL, PTB, PTC, PTDOB, PTN, PV	14
Center-right	NOVO, PR, PRB, PSC	4
Right	PP, PRTB	2
Total	**-**	**35**

Table 5: List of Brazilian political parties according to their orientation (Vasconcellos, 2016a,b).

Although some parties classify their orientation as center-left or center-right, a great portion of them can be regarded as center parties. As of 2016, the block known as "super-center" includes PEN, PHS, PP, PR, PRB, PROS, PSC, PSD, PSL, PTB, PTN and SD.

As mentioned in the introduction, the Chamber of Deputies (*Câmara dos Deputados*) is the lower house of the National Congress, comprised of 513 federal deputies (from 25 political parties), elected by a proportional representation of votes to serve a four-year term. Table 6 displays the number of elected deputies from each party/coalition, for the 2010 (2011-2014 term) and 2014 elections (2015-2018 term).

4.2 Methodology

We attempt to identify groups of deputies (and their respective parties) in the Chamber of Deputies signed networks, generated based on voting session records publicly made available by the open data initiative of the Brazilian Government [4].

To do so, we apply the $ILS - CC$ (Levorato *et al.*, 2015) procedure to solve the two problems introduced in Section II: the Correlation Clustering (CC) problem and the Symmetric Relaxed Correlation Clustering (SRCC) problem. The procedure changes the objective function that evaluates the clustering partition accordingly.

However, based on the obtained results, we chose to rely our analysis solely on SRCC clustering results [5]. The reason is that all CC solutions presented only one or two clusters as output, which, to our knowledge, did not accurately represent the political groups in the Chamber of Deputies. One possible explanation is that, as stated in Section II, when compared to the SRCC problem,

[4]The data services of the Brazilian Chamber of Deputies website can be found at http://www2.camara.leg.br/transparencia/dados-abertos

[5]We solved the SRCC problem by fixing the number of clusters (k) in the solution to $k = 4$, so as to reflect the number of coalition groups: the three main coalitions in each four-year term, listed in Tables 3 and 4, plus an additional group to represent all the candidates / parties not in a coalition.

the CC problem tends to over-evaluate the imbalance of a network, for penalizing relationships associated, for instance, with mediation processes.

Next we present several clustering results that help answering interesting questions concerning political dynamics. Each question and its respective analysis is organized in a subsection.

Remark that, whenever a clustering result is displayed as a treemap, each cluster is marked with a different color and corresponding cluster label (begins with letter C). Besides, for each cluster, the treemap displays the sum of deputies (in parenthesis), grouped by their respective party.[6] Since the clustering is based on deputies, political parties may be split into different clusters.

4.3 Evaluation of the loyalty of parties from the same coalition

We have extracted a table which, for each year, coalition and party (columns *Year*, *Party Alliance* and *Party*, respectively), gives details about the percentage of deputies from each party in each cluster (columns *C1* to *C4*). This way it is possible to spot if the majority of the deputies of a specific party does not belong to the most populous coalition cluster, which constitutes a strong evidence that such party is unfaithful to its coalition. By using this data, one can verify that, for example, in 2011 (Table 7), only 41% of PDT, 38% of PR and 42% of PRB deputies were classified inside the largest ruling coalition cluster, formed by 206 deputies. In 2012 (Table 8), only 16% (3 in 19) of PSC deputies accompanied the biggest government group, comprised of 237 deputies. Finally, in 2014 (on both network versions), just half of PT and PDT deputies followed the government coalition (see column *C1* in Table 9).

4.4 Evolution of the support of the government coalition

We start by analyzing two tables that provide, for each year and network version (columns *Year* and *Version*, respectively), the number of deputies according to their respective party alliance and the cluster to which they belong (columns *Party Alliance* and columns *C1* to *C4*, respectively). The first table (Table 10) refers to the period from 2011 to 2014 (54th legislature of the Chamber of Deputies), while the second one (Table 11) gives information about the years of 2015 and 2016 (55th legislature, corresponding to president Dilma Rousseff's second presidential term).

We observe that, in the first year of president Dilma Rousseff's government (2011), the government coalition is divided, roughly speaking, in two or three big groups, depending on the network version on which the analysis is based. According to version v1 (Figure 1), the largest cluster (C1) has 64% of the allied deputies. Also, the great majority of the president's party (PT), 82 deputies, are to be found in this cluster.

From 2012 onwards, a clear basis consolidation can be observed, with 77% of the allied deputies in the same group (cluster C1 in Figure 2). This cluster also holds more than 80 deputies of president's party (PT).

In 2013 (Figure 3), the percentage of allied deputies inside the largest cluster (C1) rises to 82% of the coalition (74 PT deputies). However, in 2014 (the last year of president Dilma Rousseff's first term), a change of course comes about. This measure falls to 66% (Figure 4) and, even worse, only about half of PT's deputies are inside the main coalition group (C3).

[6] Due to space limitation, it is not possible to show the cluster label for all groups/squares in the Figure. Please visit the website https://public.tableau.com/profile/mario.levorato to access the interactive version of the plots with full information.

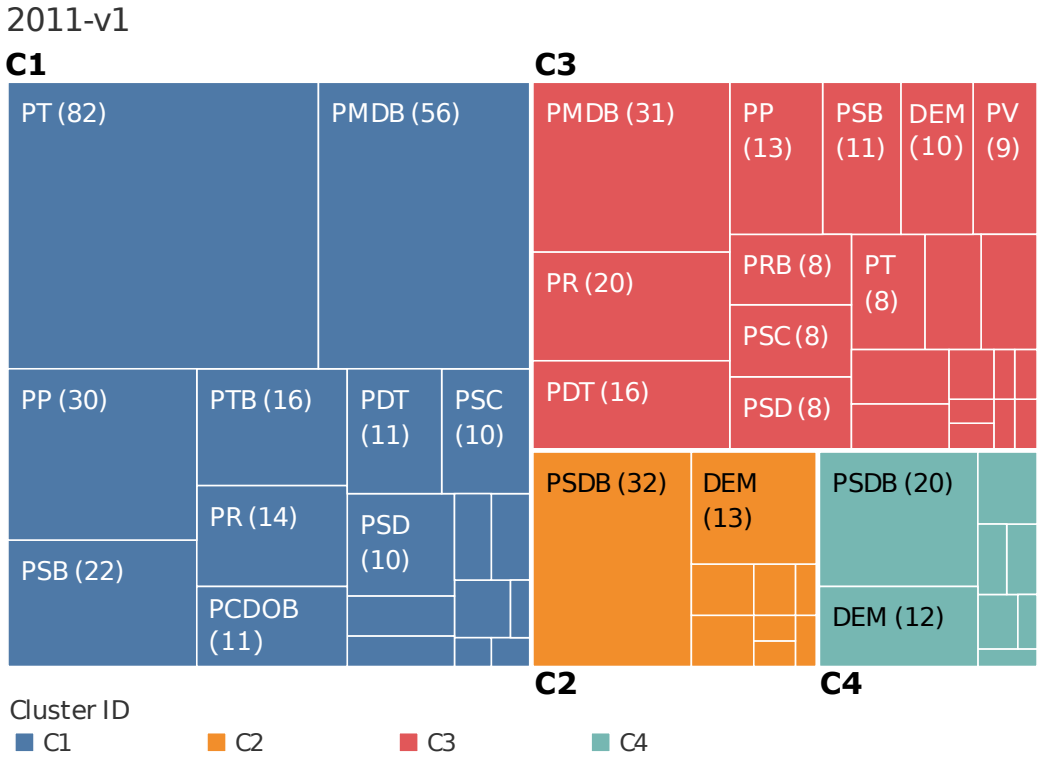


Figure 1: SRCC clustering results for the year of 2011, when solving version v1 of the voting network, by fixing the number of clusters in the solution to $k = 4$.

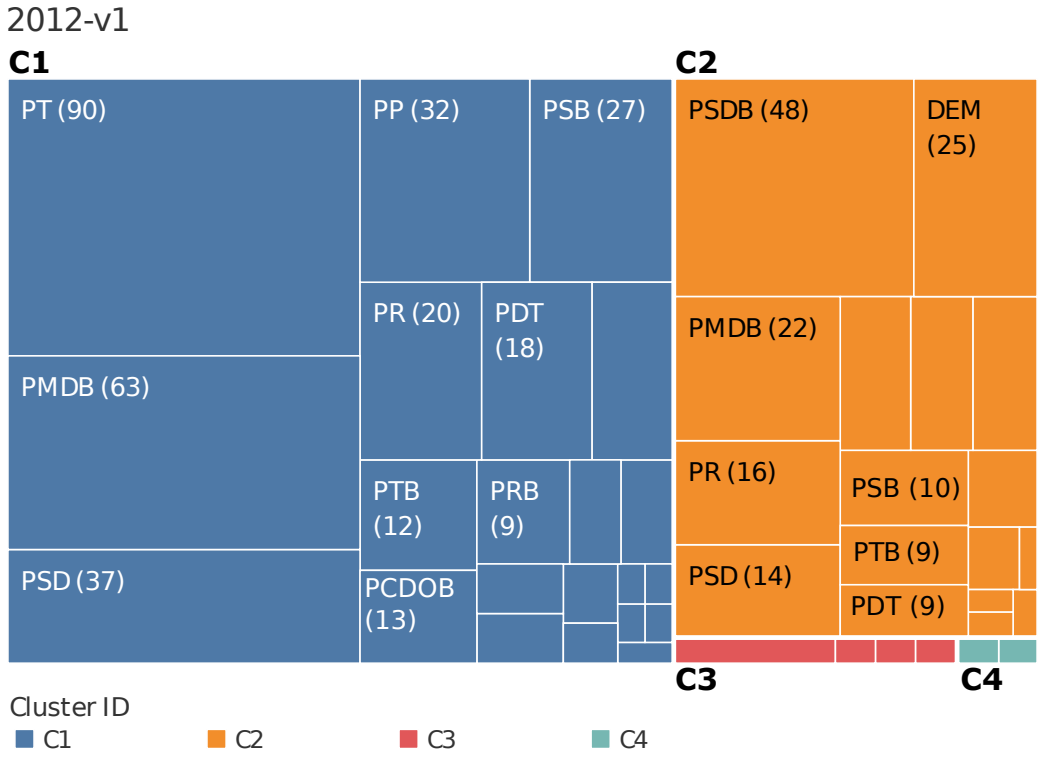


Figure 2: SRCC clustering results for the year of 2012, when using version v1 of the voting network, by fixing the number of clusters in the solution to $k = 4$.

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

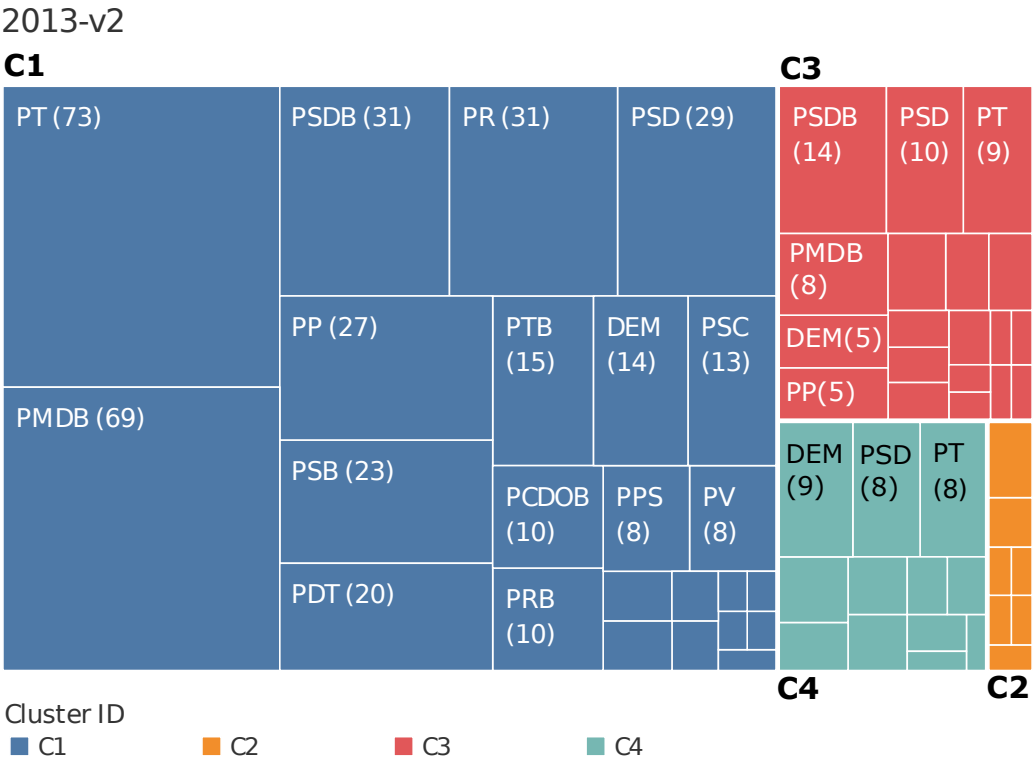


Figure 3: SRCC clustering results for the year of 2013, when using version v2 of the voting network, by fixing the number of clusters in the solution to $k = 4$.

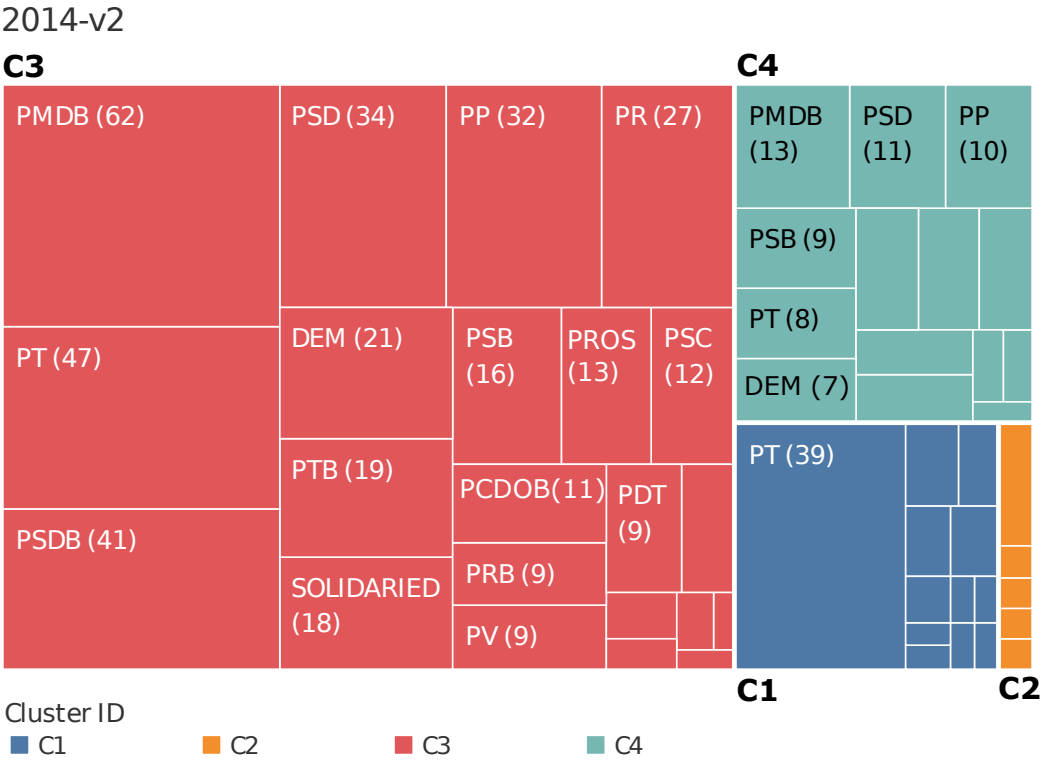


Figure 4: SRCC clustering results for the year of 2014, when using version v2 of the voting network, by fixing the number of clusters in the solution to $k = 4$.

A close look at president Dilma Rousseff's second presidential term is surprising. In 2015, the biggest group of what should be the government's new coalition (cluster C1 in Figure 5) is formed by 70% of the total number of deputies of the coalition as a whole. Notwithstanding, this group houses at most 10 deputies of the president's party (PT). Consider as well that the greatest part of PT deputies is in fact isolated in a smaller cluster, together with a few deputies from less influential parties. Note that both network versions show almost identical results[7].

A similar picture takes place in 2016 (Figure 6), when about two thirds of the supposedly allied deputies belong to the same group, which contains only 11 PT deputies. Similarly, 50 PT deputies can be found in another cluster.

Briefly speaking, results point out that in the years of 2015 and 2016, even though there are still large groups in which most deputies are from the so-called government coalition, such groups are no longer in accordance with the president's party, which is perfectly understandable because of the political crisis and the loss of parliamentary support, news widely broadcast (Dyer, 2015; Boadle, 2016; Watts, 2016a).

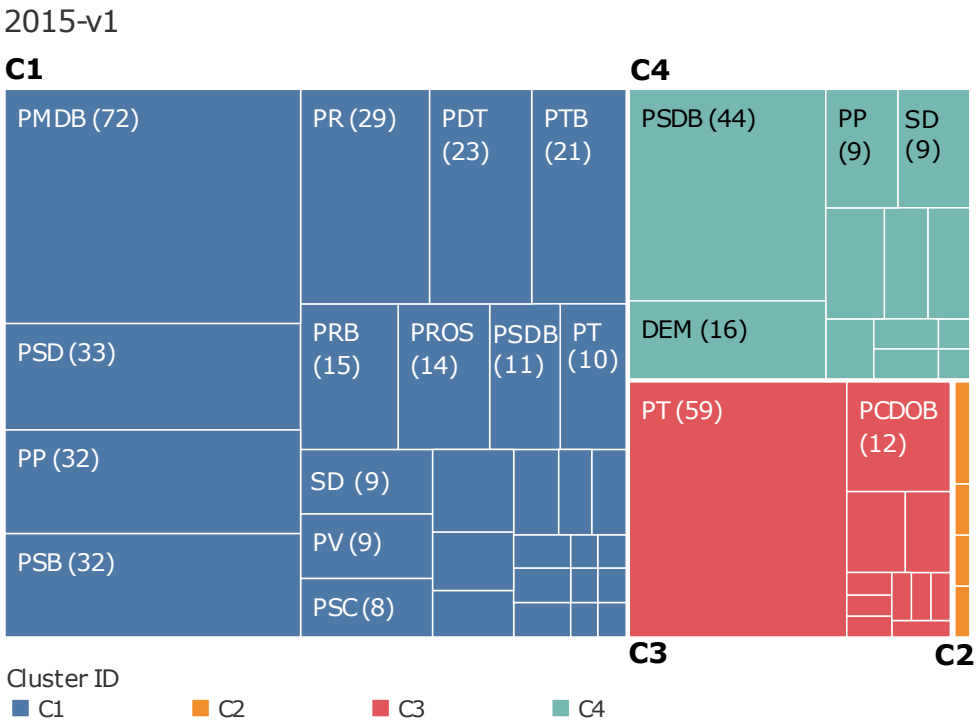


Figure 5: SRCC clustering results for the year of 2015, when using version v1 of the voting network, by fixing the number of clusters in the solution to $k = 4$.

4.5 Strength of party leadership

This study was carried out as follows: for each year from 2011 to 2016 and for each party, we scanned data about the deputies and the clusters from which they make part. This information was then cross-referenced with the cluster where the leader of the respective party is found. This way it is possible to have a clear view of how strong the leadership of each party is: if a specific deputy belongs to different cluster than its party leader, on average, this deputy did not vote the way his party expected. The full results with the information about the deputies classified

[7]Please visit https://public.tableau.com/profile/mario.levorato for a full list of charts and tables.

Volume: 2 - Year: 2017 , DOI: 10.18713/JIMIS-280217-2-3

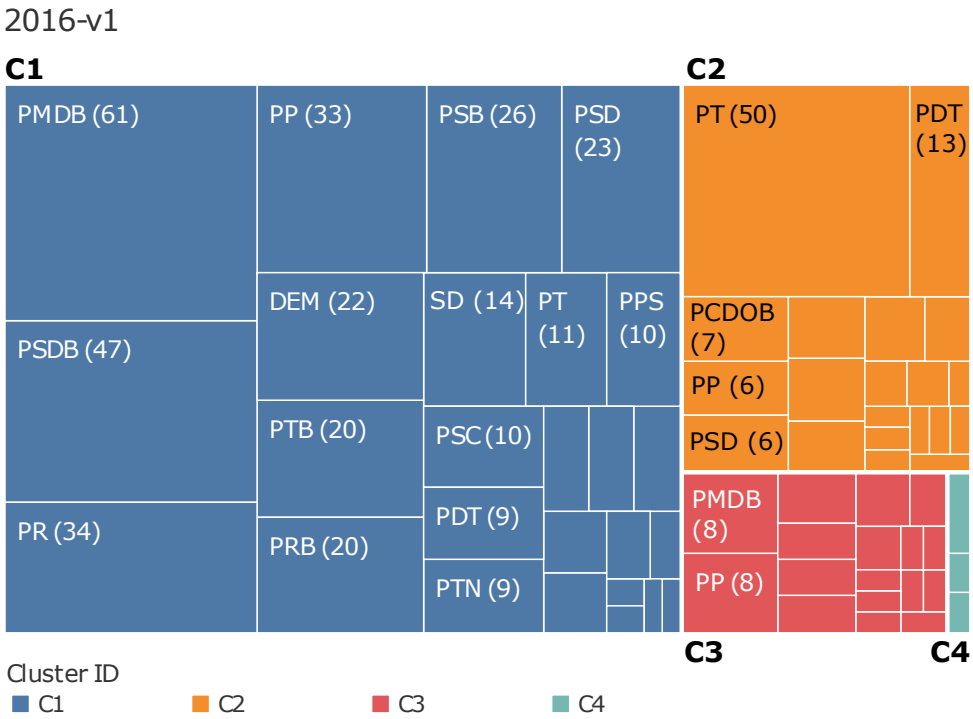


Figure 6: SRCC clustering results for the year of 2016, when using version v1 of the voting network, by fixing the number of clusters in the solution to $k = 4$.

in the same cluster as their respective party leader (percentage) are available in Table 12 for 2011-2014 and in Table 13 for 2015-2016.

For each year, the following parties have been identified as having low percentage ($\rho < 50\%$) of deputies who vote after their party leaders, independently of the analyzed network version:

- 2011: PRB, PRP, PSC, PSD;
- 2012: PR, PSB, PSC, PSD, PTB, PV;
- 2013: PCDOB, PRP, PV;
- 2014: PCDOB, PDT;
- 2015: PSC;
- 2016: PCDOB, PMB, PSL, PTDOB.

Deep consideration into this list will reveal that, as we spot a considerably great number of deputies arranged in clusters where their party leaders are not present, there is strong evidence that, on average, voting recommendations from party leaders have not been followed by many deputies.

4.6 The split of the government ruling party when the Brazilian political crisis began

The clustering results for 2014 strongly suggest that president Dilma Rousseff's party (PT) split, with 47 deputies in the first cluster, 39 in the second and 8 in the third. As seen on Figure 4, the treemap shows the fragmentation of PT in the last year of the president's first term.

4.7 Government coalition's loss of support after president Dilma Rousseff's reelection

According to the results obtained by the ILS-CC algorithm, in 2015, after president Dilma Rousseff's reelection, three clusters cover 99% of the deputies (Figure 5). The parties inside each cluster reveal the main political groups at that time:

- the largest group includes mainly center parties, such as the majority of PMDB, PSD, PP and PR;
- the second biggest group is formed by opposition parties like PSDB and DEM;
- the last one represents the government core parties, such as PT (59 deputies) and PCDOB (12 members).

A comparison between 2015 and previous years (see Figures 1, 2 and 5) reveals that the government coalition has gone through a substantial loss of support, mostly from center parties. These results anticipate a movement which became clear only the following year, when PMDB and other center-parties voted to leave the governing coalition (Boadle, 2016; Watts, 2016a; Barchfield and Savarese, 2016).

4.8 Center parties moved towards opposition when the government coalition lost power

Looking at the data for the years of 2015 (Figure 5) and 2016 (Figure 6), one can observe that the majority of center party and opposition deputies started sharing the same group. There was a strong approximation between PMDB (center), PSDB and DEM (opposition), which have previously been in separate clusters. According to the charts, one can notice that center parties have moved towards opposition.

In 2015, there was a large movement of parties from the government coalition, which went to a "super-centered" group. These parties include: PROS (12), PRB (12), PDT (22), PR (25), PP (28), PSD (33) and PMDB (71).

In 2016, the following coalition parties have effectively migrated to what can be interpreted as a huge opposition cluster: PDT(17), PRB (20), PSD (21), PP (30), PR (33) and PMDB (56).

News broadcast confirm this movement: first, the approximation between Brazil's biggest party (PMDB) and PSDB was reported (Gonçalves, 2016; Sambo and Godoy, 2016). Shortly after, PMDB voted to leave the governing alliance (Boadle, 2016; Watts, 2016a), followed by three other parties (PDT, PRB and PP) (Barchfield and Savarese, 2016).

4.9 Polarization between political groups

In 2012 (on both network versions), the chamber of deputies is polarized in two large groups (see Figure 2). The first one with 238 members, led by the majority of PT and PMDB deputies (government base). The other cluster is mainly characterized by opposition parties, such as PSDB and DEM, but it also includes dissidents from center parties like PMDB and PSD.

4.10 Relative imbalance of the analyzed signed social networks

Several authors have mentioned that real-world signed social networks are more balanced than expected (Kunegis *et al.*, 2009; Leskovec *et al.*, 2010; Kunegis *et al.*, 2010; Facchetti *et al.*, 2011). As seen on Table 14, the signed social networks generated from the Brazilian CD voting data are in fact highly balanced, which supports existing research about that topic.

V CONCLUDING REMARKS

In this article, we have investigated some of the aspects inherent to signed voting networks and political relationships, by using data from the Brazilian Chamber of Deputies (CD). We have first extracted a collection of networks based on voting patterns of the CD members. We have then applied a clustering algorithm specifically designed for signed networks, called $ILS-CC$, which aims to improve structural balance.

The clustering results allowed us to gather evidence that certain parties are indeed unfaithful to their coalition. Besides, the obtained data perfectly confirms the news broadcast about the Brazilian political situation, such as the loss of support that government coalition experienced.

Equally, the algorithm has proved to be a useful tool to spot parties under weak leadership and the existence of polarization between two large political groups. Our analysis also confirmed that the signed social networks we generated from the Brazilian CD voting data are indeed extremely balanced, hence supporting previous related works.

References

Alston L. (2016, May). *Is Dilma Rousseff's impeachment a coup or brazil's window of opportunity?. The Conversation*. Accessed: 2016-11-20. URL: https://theconversation.com/is-dilma-rousseffs-impeachment-a-coup-or-brazils-window-of-opportunity-59362.

Ames B. (1995, May). Electoral rules, constituency pressures, and pork barrel: bases of voting in the brazilian congress. *The Journal of Politics 57*(2), 324–343. doi:10.2307/2960309.

Andrade N. (2016). *House of Cunha: Who's who in the Brazilian House of Deputies*. Accessed: 2016-08-25. URL: http://houseofcunha.com.br/.

Bansal N., Blum A., Chawla S. (2002). Correlation clustering. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, Vancouver, Canada, pp. 238–250. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/sfcs.2002.1181947.

Barchfield J., Savarese M. (2016, April). *3 parties abandon brazil president as impeachment vote nears. Associated Press*. Accessed: 2016-11-20. URL: http://bigstory.ap.org/article/7995e65fd9a543a2bdc51892f02b66a3/impeachment-wracked-brazil-both-sides-gear-vote.

BBC (2016, August). *What has gone wrong in brazil? BBC*. Accessed: 2016-11-20. URL: http://www.bbc.com/news/world-latin-america-35810578.

Bevins V. (2016, March). *The politicians voting to impeach brazil's president are accused of more corruption than she is. Los Angeles Times*. Accessed: 2016-11-20. URL: http://www.latimes.com/world/mexico-americas/la-fg-brazil-impeach-20160328-story.html.

Bhattacharya A., De R. K. (2008, April). Divisive correlation clustering algorithm (dcca) for grouping of genes: detecting varying patterns in expression profiles. *bioinformatics 24*(11), 1359–1366. doi:10.1093/bioinformatics/btn133.

Boadle A. (2016, March). *Brazil's biggest party quits ruling coalition, Rousseff isolated. Reuters*. Accessed: 2016-11-20. URL: http://www.reuters.com/article/us-brazil-politics-idUSKCN0WU1AC.

Brusco M., Doreian P., Mrvar A., Steinly D. (2011). Two algorithms for relaxed structural balance partitioning: linking theory, models and data to understand social network phenomena. *Sociological Methods & Research 40*(1), 57–87. doi:10.1177/0049124110384947.

Calvão A. M., Crokidakis N., Anteneodo C. (2015). Stylized facts in brazilian vote distributions. *PLOS ONE 10*(9), e0137732. doi:10.1371/journal.pone.0137732.

Cartwright D., Harary F. (1956). Structural balance: A generalization of heider's theory. *Psychological Review 63*(5), 277–293. doi:10.1037/h0046049.

Connors W. (2016, March). *5 things to know about brazil's corruption scandal. The Wall Street Journal*. Accessed: 2016-11-20. URL: http://blogs.wsj.com/briefly/2016/03/04/5-things-to-know-about-brazils-corruption-scandal/.

Câmara d. D. (2016). *Glossário — Portal da Câmara dos Deputados (In Portuguese)*. http://www2.camara.leg.br/glossario/. Accessed: 2016-10-30. URL: http://www2.camara.leg.br/glossario/.

Dal Maso C., Pompa G., Puliga M., Riotta G., Chessa A. (2014, December). Voting behavior, coalitions and government strength through a complex network analysis. *PLoS ONE 9*(12), e116046. doi:10.1371/journal.pone.0116046.

DasGupta B., Encisob G. A., Sontag E., Zhanga Y. (2007). Algorithmic and complexity results for decompositions of biological networks into monotone subsystems. *BioSystems 90*(1), 161–178. doi:10.1016/j.biosystems.2006.08.001.

Davis J. (1967). Clustering and structural balance in graphs. *Human Relations 20*(2), 181–187. doi:10.1177/001872676702000206.

De Nooy W., Mrvar A., Batagelj V. (2011). *Exploratory social network analysis with Pajek: Revised and Expanded. 2nd Edition.*, Volume 27. Cambridge University Press. URL: http://mrvar.fdv.uni-lj.si/pajek/.

Demaine E. D., Emanuel D., Fiat A., Immorlica N. (2006). Correlation clustering in general weighted graphs. *Theoretical Computer Science 361*(2), 172–187. doi:10.1016/j.tcs.2006.05.008.

Doreian P., Mrvar A. (1996). A partitioning approach to structural balance. *Social Networks 18*(2), 149–168. doi:10.1016/0378-8733(95)00259-6.

Doreian P., Mrvar A. (2009). Partitioning signed social networks. *Social Networks 31*(1), 1–11. doi:10.1016/j.socnet.2008.08.001.

Doreian P., Mrvar A. (2015). Structural balance and signed international relations. *Journal of Social Structure 16*, 1. doi:10.1080/02664763.2015.1049517.

Drummond L., Figueiredo R., Frota Y., Levorato M. (2013). Efficient solution of the correlation clustering problem: An application to structural balance. In *Lecture Notes in Computer Science*, pp. 674–683. Springer Nature. URL: http://dx.doi.org/10.1007/978-3-642-41033-8_85.

Dyer G. (2015, August). *Frictions shake Brazil's ruling coalition. Financial Times*. Accessed: 2016-11-20. URL: https://www.ft.com/content/51d83ebe-4cd2-11e5-9b5d-89a026fda5c9.

Elsner M., Schudy W. (2009). Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, Stroudsburg, PA, USA, pp. 19–27. Association for Computational Linguistics. URL: http://dl.acm.org/citation.cfm?id=1611638.1611641.

Facchetti G., Iacono G., Altafini C. (2011). Computing global structural balance in large-scale signed social networks. In *Proceedings of the National Academy of Sciences of the United States of America*, Volume 108, pp. 20953–20958. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1109521108.

Feo T. A., Resende M. G. (1995). Greedy randomized adaptive search procedures. *Journal of Global Optimization 6*(2), 109–133. doi:10.1007/bf01096763.

Figueiredo A. C., Limongi F. (2000, January). Presidential power, legislative organization, and party behavior in brazil. *Comparative Politics 32*(2), 151–170. doi:10.2307/422395.

Figueiredo R., Moura G. (2013). Mixed integer programming formulations for clustering problems related to structural balance. *Social Networks 35*(4), 639–651. doi:10.1016/j.socnet.2013.09.002.

Gonçalves C. (2016, March). *Deputados negam que aproximação entre PSDB e PMDB seja pró-impeachment (In Portuguese). Agência Brasil*. Accessed: 2016-11-20. URL: http://agenciabrasil.ebc.com.br/politica/noticia/2016-03/deputados-negam-que-aproximacao-entre-psdb-e-pmdb-seja-pro-impeachment.

Harary F., Lim M., Wunsch D. C. (2002, July). Signed graphs for portfolio analysis in risk management. *IMA Journal of Management Mathematics 13*(3), 1–10. doi:10.1093/imaman/13.3.201.

Heider F. (1946). Attitudes and cognitive organization. *Journal of Psychology 21*(1), 107–112. doi:10.1080/00223980.1946.9917275.

Hix S. (2002, July). Parliamentary behavior with two principals: Preferences, parties, and voting in the european parliament. *American Journal of Political Science 46*(3), 688–698. doi:10.2307/3088408.

Hix S., Noury A. (2009, May). After enlargement: Voting patterns in the sixth european parliament. *Legislative Studies Quarterly 34*(2), 159–174. doi:10.3162/036298009788314282.

Huffner F., Betzler N., Niedermeier R. (2009, January). Separator-based data reduction for signed graph balancing. *Journal of Combinatorial Optimization 20*(4), 335–360. doi:10.1007/s10878-009-9212-2.

Kim S., Yoo C. D., Nowozin S., Kohli P. (2014, September). Image segmentation UsingHigher-order correlation clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*(9), 1761–1774. doi:10.1109/tpami.2014.2303095.

Kunegis J., Lommatzsch A., Bauckhage C. (2009). The slashdot zoo. In *Proceedings of the 18th international conference on World wide web - WWW '09*, pp. 741–750. Association for Computing Machinery (ACM). doi:10.1145/1526709.1526809.

Kunegis J., Schmidt S., Lommatzsch A., Lerner J., De Luca E. W., Albayrak S. (2010). Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, Volume 10, pp. 559–570. SIAM: Society for Industrial & Applied Mathematics (SIAM). doi:10.1137/1.9781611972801.49.

Leahy J. (2016, April). *Brazil's left fears Rousseff 'coup'. Financial Times*. Accessed: 2016-11-20. URL: https://www.ft.com/content/d15a8694-f621-11e5-9afe-dd2472ea263d.

Leskovec J., Huttenlocher D., Kleinberg J. (2010). Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, pp. 1361–1370. Association for Computing Machinery (ACM). doi:10.1145/1753326.1753532.

Levorato M., Drummond L., Frota Y., Figueiredo R. (2015). An ils algorithm to evaluate structural balance in signed social networks. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 1117–1122. Association for Computing Machinery (ACM). URL: http://dx.doi.org/10.1145/2695664.2695689, doi:10.1145/2695664.2695689.

Levorato M., Figueiredo R., Frota Y., Drummond L. (2017). Evaluating balancing on social networks through the efficient solution of correlation clustering problems. *EURO Journal on Computational Optimization*, 1–32. doi:10.1007/s13675-017-0082-6.

Lourenço H. R., Martin O. C., Stützle T. (2003). Iterated local search. In *Handbook of Metaheuristics*, pp. 320–353. Springer Nature. URL: http://dx.doi.org/10.1007/0-306-48056-5_11.

Macon K., Mucha P., Porter M. (2012, January). Community structure in the united nations general assembly. *Physica A: Statistical Mechanics and its Applications 391*(1-2), 343–361. doi:10.1016/j.physa.2011.06.030.

Mainwaring S., Liñán A. P. (1997). Party discipline in the brazilian constitutional congress. *Legislative studies quarterly*, 453–483. URL: http://www.jstor.org/stable/440339.

Mendonça I., Figueiredo R., Labatut V., Michelon P. (2015). Relevance of negative links in graph partitioning: A case study using votes from the european parliament. In *2015 Second European Network Intelligence Conference*, pp. 122–129. IEEE: Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/ENIC.2015.25.

Porter M. A., Mucha P. J., Newman M. E., Warmbrand C. M. (2005, May). A network analysis of committees in the u.s. house of representatives. *Proceedings of the National Academy of Sciences of the United States of America 102*(20), 7057–7062. doi:10.1073/pnas.0500191102.

Rapoza K. (2016, August). *Why the American left is mostly wrong about brazil president Dilma's impeachment. Forbes*. Accessed: 2016-11-20. URL: http://www.forbes.com/sites/kenrapoza/2016/08/26/why-brazil-presidents-impeachment-is-more-conspiracy-than-coup/#361db5c8278b.

Robins-Early N. (2016, April). *What brazil's massive corruption scandal could mean for the country. The Huffington Post*. Accessed: 2016-11-20. URL: http://www.huffingtonpost.com/entry/brazil-corruption-scandal_us_56fbf5dae4b083f5c6063e80.

Sambo P., Godoy D. (2016, March). *Brazil real, stocks gain on report prosecutors seek Lula arrest. Bloomberg*. Accessed: 2016-11-20. URL: http://www.bloomberg.com/news/articles/2016-03-10/brazil-real-extends-world-s-best-rally-as-rousseff-dealt-blow.

Shahshahani A., Nation T. (2016, August). *An international tribunal declares the impeachment of brazil's Dilma Rousseff an illegitimate coup. The Nation*. Accessed: 2016-11-20. URL: https://www.thenation.com/article/international-tribunal-declares-impeachment-of-brazils-dilma-rousseff-an-illegitimate-coup/.

Taub A. (2016, September). *All Impeachments are political. But was brazil's something more sinister? The New York Times*. Accessed: 2016-11-20. URL: http://www.nytimes.com/2016/09/01/world/americas/brazil-impeachment-coup.html.

Traag V., Bruggeman J. (2009, September). Community detection in networks with positive and negative links. *Physical Review E 80*(3), 036115. doi:10.1103/physreve.80.036115.

Vasconcellos F. (2016a). *Maioria dos partidos se posiciona como de Centro. Veja quem sobra no campo da Direita e da Esquerda — Na base dos dados - O Globo (In Portuguese)*. Accessed: 2016-10-30. URL: http://blogs.oglobo.globo.com/na-base-dos-dados/post/maioria-dos-partidos-se-posiciona-como-de-centro-veja-quem-sobra-no-campo-da-direita-e-da-esquerda.html.

Vasconcellos F. (2016b). *Partido do você não me representa (In Portuguese)*. Accessed: 2016-10-30. URL: http://infograficos.oglobo.globo.com/brasil/partido-do-voce-nao-me-representa.html.

Watts J. (2016a, March). *Brazil president closer to impeachment as coalition partner quits. The Guardian*. Accessed: 2016-11-20. URL: https://www.theguardian.com/world/2016/mar/29/brazil-president-dilma-rousseff-closer-impeachment-coalition-partner-quits.

Watts J. (2016b, September). *Brazil's Dilma Rousseff impeached by senate in crushing defeat. The Guardian*. Accessed: 2016-11-20. URL: https://www.theguardian.com/world/2016/aug/31/dilma-rousseff-impeached-president-brazilian-senate-michel-temer.

Yang B., Cheung W., Liu J. (2007, October). Community mining from signed social networks. *IEEE Transactions on Knowledge and Data Engineering 19*(10), 1333–1348. doi:10.1109/tkde.2007.1061.

Zhang Z., Cheng H., Chen W., Zhang S., Fang Q. (2008, June). Correlation clustering based on genetic algorithm for documents clustering. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pp. 3193–3198. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/cec.2008.4631230.

2010

Coalition	Party	
Dilma Rousseff (Government)	PDT	28
	PMDB	79
	PR	41
	PRB	8
	PSB	34
	PSC	17
	PT	88
	Total	**295**
Jose Serra (Opposition)	DEM	43
	PPS	12
	PSDB	54
	PTB	21
	Total	**130**
Marina Silva	PV	15
	Total	**15**
Not in a coalition	PCdoB	15
	PHS	2
	PP	41
	PSOL	3
	PTdoB	4
	Other parties	8
	Total	**73**
Total		**513**

2014

Coalition.	Party	
Dilma Rousseff (Government)	PCdoB	10
	PDT	19
	PMDB	66
	PP	37
	PR	34
	PRB	21
	PROS	11
	PSD	36
	PT	70
	Total	**304**
Aecio Neves (Opposition)	DEM	22
	PSDB	54
	PTB	25
	PTdoB	1
	SD	15
	Total	**117**
Marina Silva	PHS	5
	PPS	10
	PSB	34
	Total	**49**
Not in a coalition	PSC	13
	PSOL	5
	PV	8
	Other parties	17
	Total	**43**
Total		**513**

Table 6: Number of elected deputies from each party and coalition, for the 2010 elections (2011-2014 term, on the left) and for the 2014 elections (2015-2018 term, on the right).

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

				Cluster ID				
Year	Version	Party Alliance	Party	C1	C2	C3	C4	Total
2011	v2	Government	PDT	41.38% 12	6.90% 2	27.59% 8	24.14% 7	100.00% 29
			PMDB	67.42% 60	2.25% 2	12.36% 11	17.98% 16	100.00% 89
			PR	37.84% 14	8.11% 3	18.92% 7	35.14% 13	100.00% 37
			PRB	41.67% 5		33.33% 4	25.00% 3	100.00% 12
			PSB	64.71% 22	2.94% 1	14.71% 5	17.65% 6	100.00% 34
			PSC	61.11% 11		22.22% 4	16.67% 3	100.00% 18
			PT	91.11% 82		5.56% 5	3.33% 3	100.00% 90
			PTC				100.00% 1	100.00% 1
			Total	66.45% 206	2.58% 8	14.19% 44	16.77% 52	100.00% 310
		Opposition	DEM	7.89% 3	68.42% 26	7.89% 3	15.79% 6	100.00% 38
			PMN	75.00% 3		25.00% 1		100.00% 4
			PPS		58.33% 7	25.00% 3	16.67% 2	100.00% 12
			PSDB	1.72% 1	91.38% 53	3.45% 2	3.45% 2	100.00% 58
			PTB	72.73% 16		13.64% 3	13.64% 3	100.00% 22
			Total	17.16% 23	64.18% 86	8.96% 12	9.70% 13	100.00% 134
		Total		51.58% 229	21.17% 94	12.61% 56	14.64% 65	100.00% 444

Table 7: Party coalition and clustering details for the year of 2011, when solving version v2 of the voting network, by fixing the number of clusters in the solution to $k = 4$. For each coalition (column Party Alliance) and for each party (column Party), each cell shows the percentage of deputies of that party inside each cluster (columns C1 to C4).

				Cluster ID				
Year	Version	Party Alliance	Party	C1	C2	C3	C4	Total
2012	v2	Government	PDT	60.71% 17	3.57% 1	10.71% 3	25.00% 7	**100.00% 28**
			PMDB	80.23% 69		10.47% 9	9.30% 8	**100.00% 86**
			PR	61.11% 22		19.44% 7	19.44% 7	**100.00% 36**
			PRB	90.00% 9		10.00% 1		**100.00% 10**
			PSB	81.08% 30		13.51% 5	5.41% 2	**100.00% 37**
			PSC	15.79% 3		36.84% 7	47.37% 9	**100.00% 19**
			PT	92.47% 86	1.08% 1	4.30% 4	2.15% 2	**100.00% 93**
			PTC	100.00% 1				**100.00% 1**
			Total	**76.45% 237**	**0.65% 2**	**11.61% 36**	**11.29% 35**	**100.00% 310**
		Opposition	DEM	16.67% 5		43.33% 13	40.00% 12	**100.00% 30**
			PMN	100.00% 2				**100.00% 2**
			PPS	18.18% 2		63.64% 7	18.18% 2	**100.00% 11**
			PSDB	5.36% 3	7.14% 4	48.21% 27	39.29% 22	**100.00% 56**
			PTB	57.14% 12		14.29% 3	28.57% 6	**100.00% 21**
			Total	**20.00% 24**	**3.33% 4**	**41.67% 50**	**35.00% 42**	**100.00% 120**
		Total		**60.70% 261**	**1.40% 6**	**20.00% 86**	**17.91% 77**	**100.00% 430**

Table 8: Party coalition and clustering details for the year of 2012, when solving version v2 of the voting network, by fixing the number of clusters in the solution to $k = 4$. For each coalition (column Party Alliance) and for each party (column Party), each cell shows the percentage of deputies of that party inside each cluster (columns C1 to C4).

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

				Cluster ID				
Year	Version	Party Alliance	Party	C1	C2	C3	C4	Total
2014	v2	Government	PDT	10.53% 2	5.26% 1	47.37% 9	36.84% 7	**100.00% 19**
			PMDB	1.25% 1	5.00% 4	77.50% 62	16.25% 13	**100.00% 80**
			PR			81.82% 27	18.18% 6	**100.00% 33**
			PRB	10.00% 1		90.00% 9		**100.00% 10**
			PSB	3.85% 1		61.54% 16	34.62% 9	**100.00% 26**
			PSC		6.67% 1	80.00% 12	13.33% 2	**100.00% 15**
			PT	41.49% 39		50.00% 47	8.51% 8	**100.00% 94**
			Total	**15.88% 44**	**2.17% 6**	**65.70% 182**	**16.25% 45**	**100.00% 277**
		Opposition	DEM			75.00% 21	25.00% 7	**100.00% 28**
			PMN			66.67% 2	33.33% 1	**100.00% 3**
			PPS	11.11% 1		66.67% 6	22.22% 2	**100.00% 9**
			PSDB	2.04% 1		83.67% 41	14.29% 7	**100.00% 49**
			PTB		5.00% 1	95.00% 19		**100.00% 20**
			Total	**1.83% 2**	**0.92% 1**	**81.65% 89**	**15.60% 17**	**100.00% 109**
		Total		**11.92% 46**	**1.81% 7**	**70.21% 271**	**16.06% 62**	**100.00% 386**

Table 9: Party coalition and clustering details for the year of 2014, when solving version v2 of the voting network, by fixing the number of clusters in the solution to $k = 4$. For each coalition (column Party Alliance) and for each party (column Party), each cell shows the percentage of deputies of that party inside each cluster (columns C1 to C4).

Coalition Loyalty 2011-2014

			Cluster ID				
Year	Version	Party Alliance	C1	C2	C3	C4	Total
2011	v1	Government	199 64.19%	4 1.29%	103 33.23%	4 1.29%	310 100.00%
		Opposition	22 16.42%	48 35.82%	28 20.90%	36 26.87%	134 100.00%
	v2	Government	206 66.45%	8 2.58%	44 14.19%	52 16.77%	310 100.00%
		Opposition	23 17.16%	86 64.18%	12 8.96%	13 9.70%	134 100.00%
2012	v1	Government	238 76.77%	68 21.94%	2 0.65%	2 0.65%	310 100.00%
		Opposition	27 22.50%	89 74.17%	4 3.33%		120 100.00%
	v2	Government	237 76.45%	2 0.65%	36 11.61%	35 11.29%	310 100.00%
		Opposition	24 20.00%	4 3.33%	50 41.67%	42 35.00%	120 100.00%
2013	v1	Government	132 44.59%	2 0.68%	158 53.38%	4 1.35%	296 100.00%
		Opposition	64 57.66%		46 41.44%	1 0.90%	111 100.00%
	v2	Government	239 82.13%	6 2.06%	28 9.62%	18 6.19%	291 100.00%
		Opposition	70 65.42%	1 0.93%	22 20.56%	14 13.08%	107 100.00%
2014	v1	Government	196 70.76%	6 2.17%	41 14.80%	34 12.27%	277 100.00%
		Opposition	98 89.91%	1 0.92%	2 1.83%	8 7.34%	109 100.00%
	v2	Government	44 15.88%	6 2.17%	182 65.70%	45 16.25%	277 100.00%
		Opposition	2 1.83%	1 0.92%	89 81.65%	17 15.60%	109 100.00%

Table 10: Party coalition during the 2010 presidential elections, for the 2011-2014 term. For each year (column *Year*) and network version (column *Version*), the table shows the number of deputies in each party alliance (column *Party Alliance*) found in each cluster (columns *C1* to *C4*). Results obtained when fixing the number of clusters in the solution to $k = 4$.

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

Coalition Loyalty 2015-2016

			Cluster ID				
Year	Version	Party Alliance	C1	C2	C3	C4	Total
2015	v1	Government	229 69.60%	3 0.91%	82 24.92%	15 4.56%	329 100.00%
		Opposition	49 41.88%		3 2.56%	65 55.56%	117 100.00%
	v2	Government	210 63.83%	3 0.91%	95 28.88%	21 6.38%	329 100.00%
		Opposition	28 23.93%		7 5.98%	82 70.09%	117 100.00%
2016	v1	Government	201 63.41%	89 28.08%	23 7.26%	4 1.26%	317 100.00%
		Opposition	104 86.67%	7 5.83%	9 7.50%		120 100.00%
	v2	Government	23 7.26%	92 29.02%	186 58.68%	16 5.05%	317 100.00%
		Opposition	11 9.17%	8 6.67%	88 73.33%	13 10.84%	120 100.00%

Table 11: Party coalition during the 2014 presidential elections, for the 2015-2018 term. For each year (column *Year*) and network version (column *Version*), the table shows the number of deputies in each party alliance (column *Party Alliance*) found in each cluster (columns *C1* to *C4*). Results obtained when fixing the number of clusters in the solution to $k = 4$.

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

Party Leadership (2011-2014)

	Year / Version							
	2011		2012		2013		2014	
Party	v1	v2	v1	v2	v1	v2	v1	v2
DEM	34%	68%	83%	40%	69%	32%	89%	75%
PCDOB	73%	80%	100%	100%	23%	23%	27%	27%
PDT	55%	28%	64%	25%	44%	80%	21%	47%
PMDB	63%	67%	73%	80%	42%	82%	81%	78%
PMN					67%	67%	100%	33%
PP	67%	67%	82%	90%	31%	75%	91%	74%
PPS	33%	58%	64%	18%	55%	73%	89%	22%
PR	54%	35%	44%	19%	62%	84%	88%	82%
PRB	33%	42%	90%	90%	90%	100%	10%	90%
PROS					100%	100%	67%	62%
PRP	50%	50%			50%	50%	100%	100%
PSB			27%	14%	70%	85%	88%	62%
PSC	44%	17%	47%	47%	53%	11%	73%	80%
PSD	35%	17%	27%	18%	35%	59%	79%	71%
PSDB	34%	91%	86%	48%	65%	6%	92%	84%
PSOL	33%	67%	100%	100%	67%	67%	100%	100%
PT	91%	91%	97%	92%	62%	81%	51%	50%
PTB	73%	73%	43%	29%	68%	88%	85%	95%
PTDOB	75%	75%	100%	67%	67%	100%	100%	100%
PV	75%	33%	50%	40%	50%	20%	89%	100%

Table 12: For each party (column *Party*), displays the percentage of its deputies who vote after their party leader (i.e. deputies classified in the same group of their party leader), for each year between 2011 and 2014 (columns *2011* to *2014*) and for each network version (columns *v1* and *v2*). On certain periods, the numbers associated with a party may not have been shown. Either because the party still did not exist at that time or did not have any representation in parliament at all.

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

Party Leadership (2015-2016)

	Year / Version			
	2015		2016	
Party	v1	v2	v1	v2
DEM	30%	96%	85%	77%
PCDOB	92%	92%	50%	50%
PDT	100%	96%	59%	64%
PEN	100%	100%	75%	75%
PHS	100%	100%	100%	100%
PMB			14%	14%
PMDB	96%	95%	11%	77%
PP	78%	68%	69%	4%
PPS	58%	75%	100%	100%
PR	83%	71%	83%	80%
PRB	75%	40%	100%	100%
PROS	100%	86%	71%	71%
PRP			100%	100%
PSB	94%	68%	79%	61%
PSC	46%	43%	71%	71%
PSD	89%	89%	70%	64%
PSDB	80%	96%	87%	2%
PSL			25%	25%
PSOL	100%	100%	100%	100%
PT	83%	89%	82%	82%
PTB	81%	69%	95%	81%
PTDOB	100%	50%	50%	50%
PTN	75%	25%	82%	18%
PV	90%	20%	67%	67%
REDE	100%	100%	80%	80%

Table 13: For each party (column *Party*), displays the percentage of its deputies who vote after their party leader (i.e. deputies classified in the same group of their party leader), for the years of 2015 and 2016 (columns *2015* and *2016*; until June 2016) and for each network version (columns *v1* and *v2*). On certain periods, the numbers associated with a party may not have been shown. Either because the party still did not exist at that time or did not have any representation in parliament at all.

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

Year	**2010**		**2011**		**2012**		**2013**	
Version	v1	v2	v1	v2	v1	v2	v1	v2
$\%SRI(P)$	0.25%	0.26%	0.38%	0.43%	0.34%	0.33%	0.31%	0.35%

Year	**2014**		**2015**		**2016**	
Version	v1	v2	v1	v2	v1	v2
$\%SRI(P)$	0.07%	0.07%	0.39%	0.40%	1.92%	2.32%

Table 14: Symmetric Relaxed Imbalance ($\%SRI(P)$) measure obtained with the solution of the SRCC problem over the CD signed graphs, according to year and network version.

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-280217-2-3*

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

The Problem of Action at a Distance in Networks and the Emergence of Preferential Attachment from Triadic Closure

Jérôme KUNEGIS[*1,2], **Fariba KARIMI**[2,3], **SUN Jun**[2]

[1]University of Namur, Belgium
[2]University of Koblenz–Landau, Germany
[3]GESIS – Leibniz Institute for the Social Sciences, Germany

*Corresponding author: jerome.kunegis@unamur.be

DOI: 10.18713/JIMIS-140417-2-4
Submitted: September 6 2016 - Published: April 14 2017
Volume: **2** - Year: **2017**
Issue: **Graphs & Social Systems**
Editors: *Rosa Figueiredo & Vincent Labatut*

Abstract

In this paper, we characterise the notion of preferential attachment in networks as action at a distance, and argue that it can only be an emergent phenomenon – the actual mechanism by which networks grow always being the closing of triangles. After a review of the concepts of triangle closing and preferential attachment, we present our argument, as well as a simplified model in which preferential attachment can be derived mathematically from triangle closing. Additionally, we perform experiments on synthetic graphs to demonstrate the emergence of preferential attachment in graph growth models based only on triangle closing.

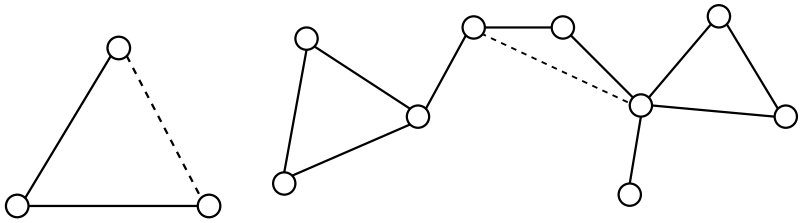
Keywords

networks; preferential attachment; triangle closing; action at a distance

I INTRODUCTION

Many natural and man-made phenomena are networks – i.e., ensembles of interconnected entities. To understand such structures is to understand their creation, their evolution and their decay. In fact, many models have been proposed for the evolution of networks, for the simple reason that a very large number of real-world systems can be modelled as networks. Rules for the evolution of networks can be broadly classified into two classes: those postulating local growth, and those postulating global growth. An example for a mechanism of local growth is triangle closing: When two people become friends because they have a common friend, then a new triangle is formed, consisting of three persons.[1] This tendency of networks to form triang-

[1] In this paper, we use the terms *triangle closing* and *triadic closure* exchangably. The notion of triadic closure



(a) Triangle closing (b) Preferential attachment

Figure 1: The two network growth mechanisms considered in this article: triangle closing and preferential attachment. In both models, new edges appear (shown as dashed lines), based on the network environment of the current graph. (a) Triangle closing: an edge is more likely to appear between nodes that have common neighbours, (b) Preferential attachment: A node with higher degree is more likely to receive an edge.

les is a natural model not only for social networks, but for almost all types of networked data. For instance, if Alice likes a movie and Bob is a friend of Alice, Bob might also come to like that movie. In this case, the triangle consists of two persons and one movie. In general, networks can contain any type of object being connected by many different types of connections, and thus many different types of such triangle closings are possible. We call this type of growth *local* because it only depends on the immediate neighbourhood of the two connected nodes; the rest of the network does not play a role.

In contrast to local graph growth rules, there is the phenomenon of preferential attachment. When, for instance, two people become friends with each other, not because they have a common friend, or go to the same class, but because one of them or both of them are popular.[2] Given a popular person, i.e. with many friends, it is more likely that he will be chosen as a friend, than an unpopular person, all else being equal. This phenomenon is referred to as preferential attachment. Preferential attachment is an often-used strategy to predict new connections, not only in social networks: a frequent movie-goer is much more likely to watch a popular film, than someone who almost never goes out to the movies watching an obscure film almost nobody knows or has seen. These types of statements seem obviously true and indeed they are used widely in application systems: recommender systems give a big preference to popular movies, search engines give higher weight to well-connected web pages, and Facebook or Twitter will make a point to show you pictures that already have many likes. In that sense, preferential attachment is true empirically, and has been verified many times in experiments. However, preferential attachment has one problematic property: It relies on connecting any two completely unrelated nodes, merely because of their degree, without considering their interconnections. Preferential attachment can thus be labelled as "action at a distance". For this reason, we argue that preferential attachment is never a primitive phenomenon, but always a derived phenomenon, emerging as a result of more basic network evolution rules, which themselves do not involve action at a distance.

So, if preferential attachment is not a primitive network evolution mechanism, which network evolution rules should then be considered as primitive in our network growth model? We will present in this paper arguments for the thesis that only the principle of triangle closing is fun-

has been alluded to multiple times in the history of the social sciences; and became mainstream with the work of Mark Granovetter (1985).

[2]Preferential attachment, too, is a concept with a long history, having been alluded to under multiple names. See the references in (Kunegis *et al.*, 2013) for an account of the early work on it.

damental, all forms of preferential attachment being derived from it. To give an argument in favour of our thesis, we will first review basic notions of networks and network evolution models, and then review preferential attachment, proposing various mechanisms by which it can arise from triangle closing, a fundamental notion in the evolution of networks. Finally, we perform experiments on synthetic graphs to test to what extent preferential attachment may emerge from graph growth models that include only triangle closing and/or random addition of edges.

II RELATED WORK

The debate over the nature of preferential attachment mechanisms dates back to the 1960s, when the economist Herbert Simon defended the role of randomness and the mathematician Benoît Mandelbrot defended the role of optimisation (Barabási, 2012). The concept of preferential attachment is also used to explain the nature of scale-free degree distributions in biological networks such as metabolic networks (Jeong *et al.*, 2000) and protein networks (Jeong *et al.*, 2001). There are various suggestions to explain the nature of preferential attachment for instance by introducing hidden variable models in which nodes possess an intrinsic fitness to other nodes in unipartite (Boguñá and Pastor-Satorras, 2003) or bipartite networks (Kitsak and Krioukov, 2011). In a recent *Nature* paper, Papadopoulos and colleagues proposed a model based on geometric optimisation of homophily space (2012). However, in these models, triadic closure is not defined as the main principle for the formation of edges.

Triadic closure, a tendency to connect to the friend of a friend (Rapoport, 1953), has been observed undeniably in many social networks such as friendship at a university (Kossinets and Watts, 2006), scientific collaborations (Newman, 2001) and in the World Wide Web (Adamic, 1999). The concept of triadic closure was first suggested by German sociologist Georg Simmel and colleagues (1950) and later on popularised by Fritz Heider and Mark Granovetter as the theory of cognitive balance in which if two individuals feel the same way about an object or a person, they seek closure by closing the triad between themselves (Heider, 2013). Since the classic preferential attachment model fails to explain the number of clusters in many social networks, many attempts have been made to include triadic closure to the model (Holme and Kim, 2002; Vázquez, 2003), in which nodes connect with certain probabilities based on the principle of triadic closure. These works have shown that the scaling law for the degree distribution and clustering coefficient can be reproduced based on these models (Klimek and Thurner, 2013). Similarly, models based on random walks as local processes have been proposed, too, of which triangle closing is a special case (Evans and Saramäki, 2005).

Hence, the scale-free nature of networks and the abundance of triangles in socio-technical networks beg for a more fundamental explanation. Moreover, the observable part of these systems is not necessarily completely representative for the entire system. Networks are generally multi-layered or multiplex, in which some layers can be hidden or simply not possible to observe (Kivelä *et al.*, 2014). For instance, the creation of a new Facebook tie can be caused by attending the same class, sharing the same hobby or living in a same neighbourhood, which is hidden from the observable data. Consequently, these *focal* points contribute to the tie creation known as *focal closure* and need to be considered in modelling realistic networks, as argued by Kossinets and Watts (2006).

III NETWORKS

The assertion that networks are to be found everywhere has become a cliché because it is true. Social networks, knowledge networks, information networks, communication networks – many papers in the field of network science motivate their use by enumerating fields in which they

play a central role. Biological networks, molecules, lexical networks, Feynman diagrams – hardly a scientific field exists in which networks do not play a fundamental role. Instead of giving a hopelessly incomplete enumeration of examples, we will simply refer the reader to the introductory section of our Handbook of Network Analysis (Kunegis, 2017), in case she wishes to convince herself of this fact. In case this is not enough, we may point to the existence of entire fields of research incorporating the word *network* and synonyms that have emerged in the last decade: network science (Börner *et al.*, 2007; Newman, 2010), web science (Hendler *et al.*, 2008), and others (Tiropanis *et al.*, 2015). There are many ways to justify the ubiquitous use of networks as a model. As an example, we may consider their use in the field of machine learning. Most classical machine learning algorithms deal with datasets consisting of data points, each consisting of the same features. Mathematically, we may model such a dataset as a set of points in a space whose dimensions are the individual features (Salton *et al.*, 1975). This formalism is very powerful, and still constitutes the backbone of many machine learning and data mining methods to this day. The standard formulation of classification, clustering and other learning problems all rely on the set-of-points-in-a-space model. However, not all machine learning problems are well described by the *set of points* model. While the set of words contained in text documents are well represented by the *bag of words* model (Baeza-Yates and Ribeiro-Neto, 1999), a social network is not. We may try to represent a social network as a bag of friends, but this representation is very unsatisfactory: each person has a set of friends, but the model does not reflect the fact that a person contained in one of these bags is the same person as one *having* a bag of friends. Thus, the vector space model cannot find connections such as "the friend of my friend" – it can only find "a person that has the same friend as me". In other words, the vector space model disconnects the role of *having friends* and that of *being a friend*. Instead, the natural way to represent friendships is as a network. Using a network model, the symmetry of the friend relationship is included automatically in the model, and relationships such as *the friend of my friend* arise as the natural way to create new edges in the network, i.e., triangle closing. In fact, we will argue that this is the only way new edges can be created in a network, and that other models are merely consequences of it, such as preferential attachment.

As an additional remark, the terms *network* and *graph* are often used interchangeably. Strictly speaking, a network is the real-world object to be analysed, such as a social network, while a graph is a mathematical structure used to model it.

IV PREFERENTIAL ATTACHMENT

Preferential attachment, also referred to by the phrase "the rich get richer", or as the Matthew effect, is observed empirically in many social networks (Kunegis *et al.*, 2013). In fact, the phenomenon of preferential attachment is known by many other names in different contexts; see the references within (Kunegis *et al.*, 2013) for an account. In other words, who has many friends, will get more new friends than who has few. Movies that have been seen by many people will be seen by more people than movies that have not. Websites that have been linked to many times will receive more new links because of this. These statements seem true, and indeed, they are true empirically for many different network types.

In fact, preferential attachment is the basis for a whole class of network models. The most basic of these, the model of Barabási and Albert (1999), describes the growth of a network, which proceeds as follow: Start with a small graph, and at each step, add a node, and connect that node to k existing nodes with a probability proportional to the number of neighbours for each existing node. In the limit where many nodes have been added in that way, the network tends to become *scale-free*, i.e. tends to have a distribution of neighbour counts that follow a power

law. Since power law degree distributions are observed in many natural networks, the usual conclusion is that preferential attachment is correct.

Preferential attachment is thus undeniably real. Why then, are we arguing against it? The reason is that preferential attachment cannot be a fundamental driving force for tie creation. How are two nodes, completely unconnected from each other, be supposed to choose to connect with each other? How can two completely disconnected nodes even *know* of each other's existence? This is a fundamental problem with all nonlocal interactions. For instance, the classical theory of gravitation as defined and used by Isaac Newton (1687) includes nonlocal interactions. In that theory, two masses exert a force on each other, regardless of their position. While the force decreases with distance, it is always nonzero, and instantaneous. The conceptual problem with this type of interaction had been identified already by Newton himself (Hesse, 1955). In modern physics, Newton's formalism is replaced by more precise theories that do not include any action at a distance. The theory of general relativity as defined by Albert Einstein (1916) for instance, only includes local interactions in the form of the Einstein field equations. Einstein's general relativity is thus free from any problematic *action at a distance*, and has been verified at many experimental scales. This is also true for other types of physical interactions – instead of a force that acts at a distance between matter particles, quantum field theory models *bosons* that connect particles. In fact, such interactions can be represented by Feynman diagrams: graph-like representations of particles in which edges are particles and nodes are interactions – any interacting particles must be connected in one diagram, directly or indirectly. In this light, we may interpret preferential attachment as a theory that is true superficially, but must be explained by an underlying phenomenon. Specifically, an underlying phenomenon that does not rely on action at a distance. As this phenomenon, we propose the known mechanism of triangle closing.

V TRIANGLE CLOSING

How do we make new friends? By meeting the friends of our friends. This represents a triangle formed by ourselves, our previous friend, and our new friend. What if we meet our new friend in another way – maybe at a party, or a concert, or at work ... in any case, there is always *some* element in common. If we meet our new friend at a party, then we are both connected to the party, and by modelling the party as a node in our network, that new friendship is indeed created by the closing of a person–person–party triangle. Of course, we may continue to ask how our connection to the party arose. After all, we did not come to a party randomly. No, we came to the party because a friend invited us, or for any other reason, as long as there is some connection. This game of connections can be played to any desired degree of precision. Maybe we *really* went from door to door until we found a party with many people. But then, how did we get from door to door? We surely must have started somewhere, likely near to our home, and have then gone on to the next door, and to the next door, and so on. In doing this, we have only followed links: We are connected to our home by living there; our home is connected to the neighbouring house, which itself is connected to the next house, and so on. This example is of course exaggerated, but serves to illustrate the principle: in order for a new edge to appear, a path has to exist from one node to another; this can go over nodes representing any type of entity, and these nodes may be visible or hidden. All in all, there is no escaping the principle of triangle closing. However we arrived at the party, it must have been by a series of triangle closings.

Thus, triangle closing fulfils the expected role as a fundamental mechanism of network growth, as it is purely local. However, we cannot deny the existence of preferential attachment, for which we must now find suitable explanations.

VI EXPLANATIONS

In recommender systems, such as those used on web sites that recommend movies to watch, preferential attachment is often taken as a solution to the cold start problem. The cold start problem in recommender systems refers to the situation in which a user has not yet entered any information about herself, and thus triangle closing cannot be used to recommend her anything. If the user has watched only a single movie, then we can find similar movies and recommend them. If a user has added only a single friend, then we can take movies liked by that friend and recommend them. But if the user is completely new, as has no friends and no ratings yet, then this strategy will not work. How then, do recommender systems give recommendations to new users? The solution is simple: they recommend the most popular items. If you subscribe to Twitter, you will be recommended popular accounts to follow. If you subscribe to Last.fm, you will be recommended popular music. For these sites, this strategy is better than not recommending anything, and in fact is a form of preferential attachment: Create, or rather recommend, links to nodes with many neighbours. How can we interpret this in terms of triangle closing? If a node has no connections yet, then surely it cannot acquire new nodes by triangle closing. How then will a node ever acquire new edges, if it starts without neighbours? The answer is that a node does not start without any neighbours. Everything is connected. A child when it is born does not start without connections; it is already connected to its parents and to its birthplace. Likewise, a user on the Web never starts from scratch: every page has a referrer, and thus the user can be connected to another website. Even if the referring web page is not known, there has to be a referrer. If a user types in a URL by hand, she has to have taken it somewhere: maybe a friend gave it to her, maybe she read it in a magazine, on a billboard, or on a truck ... in all cases, the newly created connection is not created *ex nihilo* – it is created by triangle closing.

The explanation for preferential attachment thus lies in hidden nodes: Nodes that make indirect connections between things, but do not appear in the model. On Facebook for instance, many new friendships are created between people who do not have common friends. These new friendships seemingly appear without the help of triangle closing. However, that is always due to the fact that Facebook does not know everything. Some people are simply not on Facebook, which means that if one meets a new friend through a friend that is not on Facebook and then connects the new friend via Facebook, then from the point of view of Facebook a new edge was created without triangle closing. But that is only true because Facebook does not know my initial friend. If it did, it could correctly infer the new friendship via triangle closing. Thus, any two nodes in a network can potentially be linked, even if they do not share common neighbours *in the network at hand*, because they may share a hidden common neighbour. The same argumentation applies to hidden nodes that represent non-actors, such as classes, hometowns, parties, etc.

In order to justify preferential attachment as an emergent phenomenon, we must thus derive the mechanism that leads to edges being created specifically between nodes of high degrees. Consider a network, for instance a social network. Call this the known network. Then, consider a certain number of nodes outside of that network, that are connected at random to the nodes in the known network. Call these the unknown nodes. How many common neighbours do two members of the known network have outside of the known network? Without knowing the distribution of hidden edges, this question cannot be answered. But consider that triangle closing acts not only on known–unknown–known paths, but also on known–known–unknown paths. Starting with an equal probability for all known–unknown edges, performing triangle closing will lead to the creation of known–known–unknown triangles. The newly created known–unknown

edges can then be combined with other unknown–known edges to perform, again, triangle closing, leading to new known–known edges. The result are new edges in the observed social network, with a probability proportional to the number of the initial known node's neighbours. Thus, preferential attachment emerges as a necessary consequence of iterated triangle closing, if hidden nodes are admitted. The next section will make this heuristic argument precise.

VII DERIVATION

This section gives an exemplary derivation of a simplified model that we introduce to illustrate that preferential attachment arises as a consequence of triangle closing in the presence of hidden nodes. The given scenario is very general and may be generalised easily for instance by considering multiple node types or multiple edge types. In this model, we distinguish two types of nodes: visible nodes in the set V, and hidden nodes in the set W. We will assume that there is a given, fixed number of visible nodes $|V|$, and a possibly very large number of hidden nodes $|W|$. In particular, we will consider the limit $|W| \to \infty$.

Let $G = (V \cup W, E)$ be the graph representing the complete system, in which V is the set of visible nodes, and W the set of hidden nodes. Additionally, E is the set of edges connecting nodes in V with nodes in W. While we assume that the individual edges in E are hidden, the degree of the nodes in V is not hidden. In other words, the number of edges of E incident to each node in V is known. Edges between nodes in V will not be considered. Likewise, edges between nodes in W need not be considered, since they do not contribute to the degree of nodes in V. Thus, the considered network G is bipartite. We will use the convention that $n = |W|$, and the degree of a node u is denoted by $d(u)$. We now assume that the graph G will receive new edges according to the principle of triangle closing. Thus, two nodes in V will connect with a probability proportional to the number of common neighbours they have. Seeing only nodes in V and their degree, preferential attachment can then be observed as described in the following.

In order to make our derivation, we need to make two assumptions:

- The triangle closing process is random in the sense that new edges are added between any possible node pairs with equal probability.
- The typical degree of nodes is significantly smaller than the number of nodes, i.e., $d(x) \ll n$. This is precise when n goes to infinity.

Let $u, v \in V$ be two nodes of the network. Under the assumption that the edges are distributed randomly in the graph, the probability p that u and v are connected can be derived combinatorically by considering the number of configurations in which the two nodes do not share a common neighbour. Given that u and v have degree $d(u)$ and $d(v)$ respectively, the total number of configurations for the edges connected to the nodes is

$$\binom{n}{d(u)}\binom{n}{d(v)}. \tag{1}$$

Out of those, the number of configurations in which the neighbours of the two nodes are disjoint is given by

$$\binom{n}{d(u)}\binom{n-d(u)}{d(v)}. \tag{2}$$

Thus, the probability that the two nodes share a common neighbour is given by

$$p = 1 - \frac{\binom{n}{d(u)}\binom{n-d(u)}{d(v)}}{\binom{n}{d(u)}\binom{n}{d(v)}} = 1 - \frac{\binom{n-d(u)}{d(v)}}{\binom{n}{d(v)}}. \tag{3}$$

We now use the falling factorial to express binomial coefficients, i.e.,

$$n^{\underline{a}} = n(n-1)(n-2)\cdots(n-a+1). \tag{4}$$

The falling factorial has the property that in the limit where a is constant and n goes to infinity, we have

$$\lim_{n\to\infty} \frac{n^{\underline{a}}}{n^a} = 1 \tag{5}$$

and also,

$$\binom{n}{a} = \frac{n^{\underline{a}}}{a!}, \tag{6}$$

and thus

$$p = 1 - \frac{(n-d(u))^{\underline{d(v)}}d(v)!}{d(v)!n^{\underline{d(v)}}} = 1 - \frac{(n-d(u))^{\underline{d(v)}}}{n^{\underline{d(v)}}}. \tag{7}$$

In the limit when n goes to infinity we may thus assume that

$$p = 1 - \frac{(n-d(u))^{d(v)}}{n^{d(v)}} = 1 - \left(1 - \frac{d(u)}{n}\right)^{d(v)} \tag{8}$$

and using again the limit $n \to \infty$, and the property that in the limit where ε goes to zero, $(1-\varepsilon)^k$ goes to $(1-k\varepsilon)$,

$$p = \frac{d(u)d(v)}{n}. \tag{9}$$

It thus follows that $p \sim d(u)d(v)$, i.e., the probability of the nodes u and v being connected is proportional to both $d(u)$ and $d(v)$. Thus, we find that preferential attachment is a consequence of the triangle closing model. Preferential attachment itself then leads to a scale-free degree distribution, as per Barabási and Albert (1999).

VIII EXPERIMENTS

In this section, we give empirical evidence for the emergence of preferential attachment in graph growth models that do not include it. In the experiments, we generate synthetic networks via a random growth process that does not include preferential attachment, as well as using random growth processes that do include preferential attachment. In all generated networks, effects of preferential attachment are then measured empirically. All generated networks have 1,000 nodes and 10,000 edges, and are undirected, loopless, and do not allow multiple edges. In all cases, the graphs are generated by starting with a graph of 1,000 nodes and without edges, and adding edges one by one. For each edge that is added, one of the following three methods is chosen at random:

- **Random**: With probability p_{r}, an edge is added randomly between two unconnected nodes. All pairs of distinct unconnected nodes are chosen with equal probability.
- **Triangle closing**: With probability p_{tc}, among all unclosed triads, one is chosen randomly with equal probability, and the third edge is added. An unclosed triad is a triple of nodes (u, v, w) such that (u, v) and (u, w) are connected, but v and w are not connected. If chosen, the triangle is completed by adding the edge (v, w). If no unclosed triads are present, an edge is added at random as described in the previous case.
- **Preferential attachment**: With probability p_{pa}, a node is chosen with a probability proportional to the node's current degree. Then, out of all nodes not connected to that node, one is chosen randomly and with equal probability, and an edge is added between the two selected nodes. If there is not at least one unconnected pair of nodes with nonzero degree, an edge is added at random as described in the first case.

In each experimental trial, the three probabilities are chosen such that $p_{\mathrm{r}} + p_{\mathrm{tc}} + p_{\mathrm{pa}} = 1$. Each of these probabilities is varied from 0 to 1 in increments of 1/11, excluding the case $p_{\mathrm{r}} = 0$ in order to avoid the runaway case of an individual node accumulating all edges.[3]

First, in order to verify whether a graph created by the process of triangle closing display scale-free behaviour, we compare the generated distribution of the triangle closing case with the degree distributions for the random and preferential attachment cases.[4] All three degree distributions are shown in Figure 2. In the plot, several observations can be made. The degree distribution for the triangle closing case displays power law-like behaviour over multiple orders of magnitude, from the smallest degrees of one, to approximately one hundred. While the networks generated by triangle closing and preferential attachment have similar power-like degree distribution, both with comparable exponent, we must note that the maximum degree in the preferential attachment case is larger than in the triangle closing case. However, the triangle closing model displays a power law degree distribution with exponential cut-off that has been observed in many real networks due to finite size effect (Boguñá *et al.*, 2004; Clauset *et al.*, 2009). For comparison, the preferential attachment case also displays power law-like behaviour, although not for very small degrees (under about 10), and additionally has a well-defined long tail. The purely random case leads to a degree distribution that shows no scale-free behaviour.

We measure the equality of the distribution of edges, or its opposite, its skewness, as the primary consequence of the preferential attachment process. As a measure, we use the Gini coefficient of the degree distribution, as defined in (Kunegis and Preusse, 2012). The Gini coefficient is zero when all nodes have equal degree, and attains its theoretical maximum of one when all nodes except a single one have degree zero.[5]

The experimental results are shown in Figure 3. In the triangle shown in the figure, the top-to-bottom-right edge shows the cases in which preferential attachment is excluded, while the bottom-left corner (*Pref. att. = 100%*) represents the case of exclusive preferential attachment. As expected, the 100% random case results in an Erdős–Rényi graph in which the degrees have a Poisson distribution, and thus a very uniform number of edges over all nodes, giving a small Gini coefficient of 17.7%. The pure preferential attachment case gives a higher value

[3] In the degenerate case of $p_{\mathrm{r}} = 0$, almost all edges will be attached to a single node in the $n \to \infty$ limit.

[4] As described in the previous paragraph, the cases of pure triangle closing and preferential attachment also include 1/11 of edges based on random assignment.

[5] Since an edge always connects two nodes, the actual maximum is attained in star graphs, in which all edges attach to a single node, and other nodes have a degree of zero and one. In the large-graph limit, the Gini coefficient in such graphs tends to one.

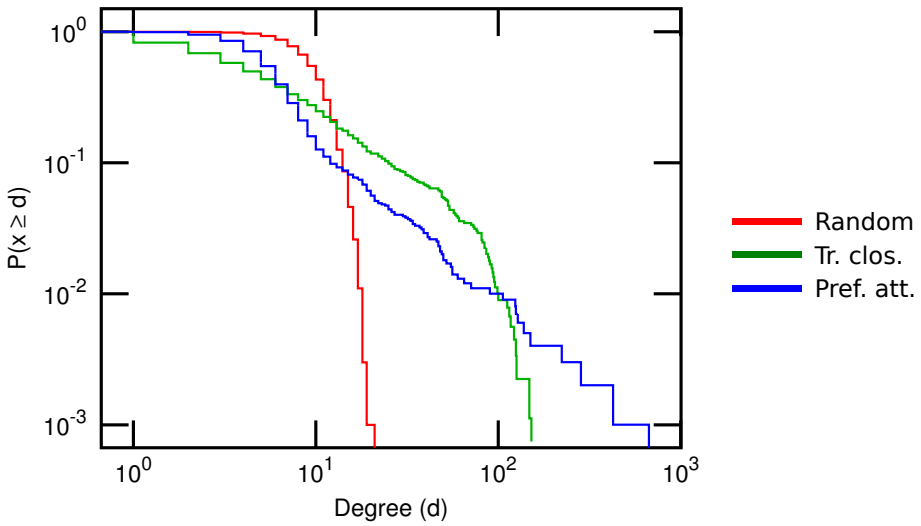


Figure 2: The cumulative degree distribution for the three extremal generated networks.

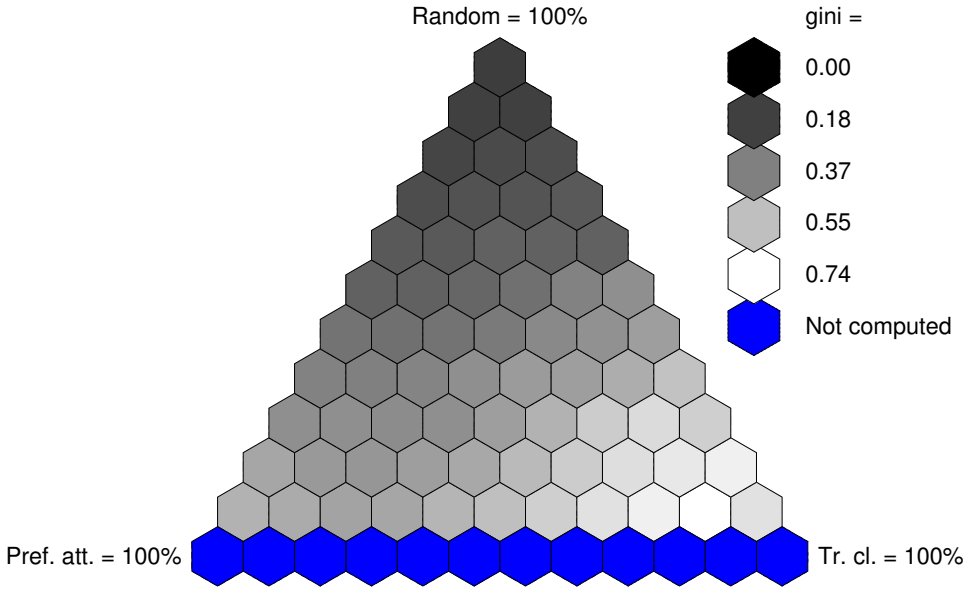


Figure 3: Experimental results: Each cell shows one experimental run with a different probability of adding each edge at random (top), via triangle closing (bottom right), and preferential attachment (bottom left). The bottom row was not executed due to the tendency of models with random edges to attach all edges to a single node, giving values of the Gini coefficient very close to the theoretical maximum of one.

Volume: **2** - Year: **2017** , *DOI: 10.18713/JIMIS-140417-2-4*

of about 51.5%.[6] The pure triangle closing method results in a value of the Gini coefficient of 65.1%, a value similar to (and even superior to) the value in the pure preferential attachment case. Thus, it is indeed the case that a skewed degree distribution is generated by a purely local process of triangle closing, without the need for explicit preferential attachment. We note also that preferential attachment is observed even though the number of nodes in the network ($n = 1{,}000$) is relatively small when compared to the theoretical model described in the previous section in which the limit $n \to \infty$ is taken.

IX DISCUSSION

Our experiments have allowed us to observed that triangle closing leads to skewed and scale-free degree distributions. However, the status of a mechanism as *fundamental* is not clear cut. When a phenomenon is explained by another, more fundamental phenomenon, we can consider it as derived. But how can we be sure that a phenomenon is not explained by a more basic phenomenon? What does it mean for a phenomenon to be fundamental? Just as physics cannot declare one theory to be final, we cannot declare one network growth mechanism to be final. Thus, individual instances of triangle closing can for instance be explained by several layers of triangle closing, just as in physics a direct interaction can be explained by a new mediating particle. In the end however, this applies only to specific instances of triangle closing, as it replaces them with other, more detailed instances of triangle closing. Thus, triangle closing *does* play a fundamental role in growing network models, only that it cannot always be derived which three nodes are taking part in it, as one of the three nodes is often hidden. In the end, the only judge of the validity of a model remains the experiment, and in practice, used models do not have to be fundamental – recommenders and information retrieval systems have had enough success by applying preferential attachment directly.

As mentioned in the introduction, triangle closing is itself a general phenomenon that not only applies to pure social networks, but also to other types of networks. In the case of property networks, i.e., networks containing edges between persons and the properties they have, triangle closing can be identified with the concept of homophily, i.e., the concept that friends tend to be similar. As an example, the fact that two smokers become friends can be modelled as the closure of the (person A)–(colleague + smoker)–(person B) triangle, in which "colleague + smoker" is a non-person node of the network representing the property of *being a colleague and a smoker*. Thus, the fact that friends of smokers are more likely to be smokers too (a classical example of homophily) can be analysed as a form of triangle closing in a graph that is not purely a social network, as it contains non-person nodes. Homophily is thus consistent with the view that triangle closing is fundamental (Shalizi and Thomas, 2011).

The problem posed in this paper can be generalised to other graph growth mechanisms. For instance, we may ask whether assortativity (the tendency of connected nodes to have correlated degrees) or community structures emerge from triangle closing alone. In the case of community structures, triangle closing trivially plays a role, as triangle closing by construction leads to tightly connected graphs. As for assortativity, the fact that both assortativity (a positive correlation between degrees) and dissortativity (a negative correlation between degrees) have been observed in social networks points to the fact that a single model such as triangle closing cannot (and is not expected to) explain all properties of a social network, and other phenomena must be at work, which may or may not be local.

[6] In this and all subsequent cases labelled as *pure*, the method in question has a probability of $p_{\mathrm{tc,pa}} = 10/11$ while a random edge is added with a probability of $p_{\mathrm{r}} = 1/11$.

References

Adamic L. A. (1999). The small world Web. In *International Conference on Theory and Practice of Digital Libraries*, pp. 443–452. doi:10.1007/3-540-48155-9_27.

Baeza-Yates R., Ribeiro-Neto B. (1999). *Modern Information Retrieval*. Boston, USA: Addison-Wesley.

Barabási A.-L. (2012). Network science: Luck or reason. *Nature 489*(7417), 507–508. doi:1038/nature11486.

Barabási A.-L., Albert R. (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512. doi:10.1126/science.286.5439.509.

Boguñá M., Pastor-Satorras R. (2003). Class of correlated random networks with hidden variables. *Physical Review E 68*(3), 036112. doi:10.1103/PhysRevE.68.036112.

Boguñá M., Pastor-Satorras R., Vespignani A. (2004). Cut-offs and finite size effects in scale-free networks. *European Physical Journal B 38*(2), 205–209. doi:10.1140/epjb/e2004-00038-8.

Börner K., Sanyal S., Vespignani A. (2007). Network science. *Annual Review of Information Science and Technology 41*(1), 537–607. doi:10.1002/aris.2007.1440410119.

Clauset A., Shalizi C. R., Newman M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review 51*(4), 661–703. doi:10.1137/070710111.

Einstein A. (1916). Die Grundlagen der allgemeinen Relativitätstheorie. *Annalen der Physik 49*, 769–822. doi:10.1002/andp.19163540702.

Evans T. S., Saramäki J. P. (2005). Scale-free networks from self-organization. *Physical Review E 72*(2), 026138. doi:10.1103/PhysRevE.72.026138.

Granovetter M. (1985). The strength of weak ties. *American Journal of Sociology 91*, 481–510. doi:10.1086/225469.

Heider F. (2013). *The Psychology of Interpersonal Relations*. Abingdon, UK: Psychology Press.

Hendler J., Shadbolt N., Hall W., Berners-Lee T., Weitzner D. (2008). Web science: An interdisciplinary approach to understanding the Web. *Communications of the ACM 51*(7), 60–69. doi:10.1145/1364782.1364798.

Hesse M. B. (1955). Action at a distance in classical physics. *Isis 46*(4), 337–353. doi:10.1086/348429.

Holme P., Kim B. J. (2002). Growing scale-free networks with tunable clustering. *Physical Review E 65*(2), 026107. doi:10.1103/PhysRevE.65.026107.

Jeong H., Mason S. P., Barabási A.-L., Oltvai Z. N. (2001). Lethality and centrality in protein networks. *Nature 411*(6833), 41–42. doi:10.1038/35075138.

Jeong H., Tombor B., Albert R., Oltvai Z. N., Barabási A.-L. (2000). The large-scale organization of metabolic networks. *Nature 407*(6804), 651–654. doi:10.1038/35036627.

Kitsak M., Krioukov D. (2011). Hidden variables in bipartite networks. *Physical Review E 84*, 026114. doi:10.1103/PhysRevE.84.026114.

Kivelä M., Arenas A., Barthelemy M., Gleeson J. P., Moreno Y., Porter M. A. (2014). Multilayer networks. *Journal of Complex Networks 2*(3), 203–271. doi:10.1093/comnet/cnu016.

Klimek P., Thurner S. (2013). Triadic closure dynamics drives scaling laws in social multiplex networks. *New Journal of Physics 15*(6), 063008. doi:10.1088/1367-2630/15/6/063008.

Kossinets G., Watts D. J. (2006). Empirical analysis of an evolving social network. *Science 311*(5757), 88–90. doi:10.1126/science.1116869.

Kunegis J. (2017). *Handbook of Network Analysis [KONECT – the Koblenz Network Collection]*. URL: `http://konect.uni-koblenz.de/downloads/konect-handbook.pdf`.

Kunegis J., Blattner M., Moser C. (2013). Preferential attachment in online networks: Measurement and explanations. In *5th Annual ACM Web Science Conference*, pp. 205–214. doi:10.1145/2464464.2464514.

Kunegis J., Preusse J. (2012). Fairness on the web: Alternatives to the power law. In *4th Annual ACM Web Science Conference*, pp. 175–184. doi:10.1145/2380718.2380741.

Newman M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E 64*(2), 025102. doi:10.1103/PhysRevE.64.025102.

Newman M. E. J. (2010). *Networks: An introduction*. Oxford, UK: Oxford University Press.

Newton I. (1687). *Philosophiæ Naturalis Principia Mathematica*. Oxford, UK: Edmund Halley. First edition.

Papadopoulos F., Kitsak M., Serrano M. A., Boguñá M., Krioukov D. (2012). Popularity versus similarity in growing networks. *Nature 489*(7417), 537–540. doi:10.1038/nature11459.

Rapoport A. (1953). Spread of information through a population with socio-structural bias: II. Various models with partial transitivity. *The bulletin of mathematical biophysics 15*(4), 535–546. doi:10.1007/BF02476441.

Salton G., Wong A., Yang C. S. (1975, November). A vector space model for automatic indexing. *Communications of the ACM 18*(11), 613–620. doi:10.1145/361219.361220.

Shalizi C. R., Thomas A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research 40*(2), 211–239. doi:10.1177/0049124111404820.

Simmel G., Wolff K. H. (1950). *The Sociology of Georg Simmel*. New York City, USA: Simon and Schuster.

Tiropanis T., Hall W., Crowcroft J., Contractor N., Tassiulas L. (2015). Network science, Web science, and internet science. *Communications of the ACM 58*(8), 76–82. doi:10.1145/2699416.

Vázquez A. (2003). Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E 67*(5), 056104. doi:10.1103/PhysRevE.67.056104.

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

Analyse de réseaux criminels de traite des êtres humains: méthodologie, modélisation et visualisation

Bénédicte LAVAUD-LEGENDRE[1,2], **Cécile PLESSARD**[3], **Antoine LAUMOND**[1,4], **Guy MELANÇON**[*1,4], **Bruno PINAUD**[1,4]

[1]Université de Bordeaux
[2]CNRS UMR 5114 COMPTRASEC, France
[3]CNRS UMR 5193 LISST, Toulouse, France
[4]CNRS UMR 5800 LaBRI, France

*Correspondance : guy.melancon@u-bordeaux.fr

DOI : 10.18713/JIMIS-300617-2-5
Soumis le 8 octobre 2016 - Publié le 18 juillet 2017
Volume : **2** - Année : **2017**
Titre du numéro : **Graphes & systèmes sociaux**
Éditeurs : *Rosa Figueiredo & Vincent Labatut*

Résumé

Cet article dessine le contexte d'une étude portant sur les réseaux criminels de traite des êtres humains et décrit la rencontre de trois champs disciplinaires engagés dans ces travaux : Droit, Sociologie et Informatique, ainsi que les éléments méthodologiques développés. Il pose les fondations d'une méthodologie venant en appui à l'étude juridique des réseaux criminels, et plus spécifiquement de ceux se livrant à des faits de traite des êtres humains. La "science des réseaux" (Network Science), vue à la fois comme une abstraction mathématique et une approche et méthodologie sociologique, sert de socle pour formuler et explorer un faisceau d'hypothèses éclairant le(s) mode(s) opératoire(s) des réseaux criminels. Les leçons apprises, nourries des interactions entre disciplines, permettent de dessiner les axes de travaux futurs pour améliorer la méthodologie avancée.

Mots-Clés

réseaux criminels ; traite des êtres humains ; exploitation ; données judiciaires ; analyse de réseaux

I INTRODUCTION

Depuis la fin des années 80, la combinaison entre l'attractivité des pays d'europe de l'Ouest d'une part, et les difficultés d'accès à une migration légale dans ces mêmes pays d'autre part, encouragent le développement de pratiques criminelles permettant à ceux qui veulent quitter leur pays de franchir les frontières, au prix, dans certains cas, de leur soumission à des faits

d'exploitation. Or ces pratiques violent un certain nombre de valeurs fondamentales de nos sociétés : respect des droits de l'homme, protection des frontières, règles de la concurrence...

Aussi, à compter des années 90, la répression de ces faits a constitué pour bon nombre d'Etats un défi considérable, justifiant une importante mobilisation politique et l'adoption de normes juridiques internationales [1] et nationales [2]. Le Protocole des Nations-Unies, dit Protocole de Palerme vise à sanctionner les actes qui préparent l'exploitation d'une personne. Plus précisément, il s'agit "dans le but de son exploitation", de "recruter, transporter, transférer, héberger, accueillir une personne", par "enlèvement, fraude, tromperie, abus d'autorité ou d'une situation de vulnérabilité, ou par l'offre ou l'acceptation de paiements ou d'avantages pour obtenir le consentement d'une personne ayant autorité sur une autre" [3] .

Or, la notion d'exploitation n'est définie ni en droit interne [4] ni en droit international [5]. Elle n'est qualifiée que par référence à des pratiques : exploitation de la prostitution d'autrui, exploitation sexuelle, travail ou services forcés, esclavage ou pratiques analogues à l'esclavage, servitude ou prélèvement d'organe, etc.

Ce constat est à mettre en parallèle avec les difficultés rencontrées dans la répression de cette activité criminelle. Dans son Rapport sur les progrès réalisés dans la lute contre la traite des êtres humains en 2016, la Commission européenne indique : "Le taux de poursuites et de condamnations reste faible, ce qui est inquiétant, surtout si on le compare au nombre de victimes identifiées." [6]. Aussi, nous posons le postulat qu'au-delà d'éléments contextuels [7], ces difficultés ont précisément pour origine l'absence de définition notionnelle de l'exploitation précédemment évoquée. C'est sur ce postulat que repose notre travail de recherche et la démarche scientifique entreprise : observer les pratiques criminelles afin d'élaborer une définition de l'exploitation.

On rappellera que l'élaboration d'une loi procède de différentes étapes : identification de la nécessité de légiférer (définition du problème et de ses causes réelles) ; détermination des buts et

1. Protocole additionnel à la Convention des Nations-Unies de lutte contre la traite des êtres humains, Palerme, 15 novembre 2000, Recueil des traités, vol. 2225, no 39574 ; Convention du Conseil de l'Europe consacrée à la lutte contre la traite des êtres humains, Varsovie, 16 mai 2005, Série des Traités du Conseil de l'Europe, no 197 ; directive 2004/81/CE du 29 avril 2004 ; directive 2011/36/UE du 5 avril 2011 (ayant remplacé la décision-cadre 2002/629/JAI du 19/07/2002).

2. Loi 2003-239 du 18/03/2003, 2013-711 du 5/08/2013, Décret 2007-1352 du 13/09/2007, loi 2013-711 du 5/08/2013, loi 2016-444 du 13/04/2016.

3. Protocole de Palerme, préc., article 3 a) : "Par enlèvement, fraude, tromperie, abus d'autorité ou d'une situation de vulnérabilité, ou par l'offre ou l'acceptation de paiements ou d'avantages pour obtenir le consentement d'une personne ayant autorité sur une autre en vue de l'exploitation".

4. Article 225-4-1 du Code pénal : "L'exploitation mentionnée au premier alinéa du présent I est le fait de mettre la victime à sa disposition ou à la disposition d'un tiers, même non identifié, afin soit de permettre la commission contre la victime des infractions de proxénétisme, d'agression ou d'atteintes sexuelles, de réduction en esclavage, de soumission à du travail ou à des services forcés, de réduction en servitude, de prélèvement de l'un de ses organes, d'exploitation de la mendicité, de conditions de travail ou d'hébergement contraires à sa dignité, soit de contraindre la victime à commettre tout crime ou délit".

5. L'article 3a du Protocole de Palerme indique sur ce point : "L'exploitation comprend, au minimum, l'exploitation de la prostitution d'autrui ou d'autres formes d'exploitation sexuelle, le travail ou les services forcés, l'esclavage ou les pratiques analogues à l'esclavage, la servitude ou le prélèvement d'organes".

6. Rapport de la Commission au Parlement européen et au Conseil, Rapport sur les progrès réalisés dans la lutte contre la traite des êtres humains (2016), Bruxelles 19 mai 2016 COM(2016) 267 final.

7. L'origine immédiate de ces difficultés renvoie au caractère récent de ces pratiques, à la complexité de l'infraction, à la diversité des champs de politiques publiques dont elles relèvent ainsi qu'à la difficulté des victimes à dire les faits subis. Sur ces différents points, voir : CNCDH, La lutte contre la traite et l'exploitation des êtres humains, Rapport 2015, La Documentation française, 2016.

objectifs de la loi ; élaboration de moyens susceptibles d'atteindre les objectifs visés (élaboration dans notre contexte de l'incrimination de traite) (Delley et Flueckiger, 2005). Aussi, en s'appuyant conjointement sur cette démarche juridique et sur une approche empirique propre à la sociologie, il s'agit de revenir très précisément à l'origine du problème, en observant concrètement les pratiques criminelles. Cette recherche a donc pour objectif de préciser la substance de l'exploitation en observant le contexte dans lequel elle s'exerce ainsi que la relation en elle-même. Elle suit ainsi la logique inhérente à l'infraction de traite des êtres humains qui sanctionne les actes qui rendent possible l'exploitation, soit son contexte. Si les réticences des victimes à témoigner sont à rapprocher de la mise en place de mécanismes d'emprise au sens psychologique du terme (Lavaud-Legendre et Quattoni, 2013), dans le cadre d'une analyse sociologique de la pratique criminelle, on peut supposer que ces réticences sont également le résultat de stratégies destinées à isoler les victimes de leur environnement et à les rendre dépendantes de ceux qui tirent profit de leur activité. Autrement dit, nous posons ici comme hypothèse initiale que ces stratégies participent de l'exploitation et, de ce fait, en éclairent la substance.

Cette approche implique de trouver un matériau de recherche permettant d'observer "à la source" les pratiques criminelles. Au regard des objectifs visés, le choix de travailler sur un dossier judiciaire s'est imposé.

Un tel dossier rassemble l'ensemble des pièces constituées au cours des phases de l'enquête, puis de l'instruction pénale, en vue du renvoi des personnes impliquées dans des faits pénalement répréhensibles par devant les juridictions de jugement [8]. Une convention a donc été conclue avec un tribunal d'une ville de province pour obtenir la remise d'un tel dossier, soit quelques 50 000 pages. En l'occurrence, il s'agissait d'un dossier ayant donné lieu à un jugement définitif et sanctionnant des faits de traite des êtres humains à des fins d'exploitation sexuelle commis par des ressortissants nigérians.

La lecture de ce premier dossier associée à notre connaissance de ces pratiques criminelles, nous a permis de formuler notre hypothèse de recherche que nous présenterons avant le contexte scientifique dans lequel s'inscrit ce projet et sa dimension pluridisciplinaire.

1.1 Hypothèse de recherche

Dès lors, notre hypothèse principale est que l'exploitation, indépendamment du type d'activité (prostitution, esclavage domestique, travail forcé dans les travaux publics, la restauration, etc.) et du degré de contrainte, ne peut s'exercer que dans un contexte relationnel spécifique. Ce contexte se caractériserait par une organisation maîtrisée de l'ensemble des relations entre les individus prenant part à une activité criminelle quel que soit leur rôle et position au sein de celle-ci. La dimension maîtrisée de ces relations sera principalement questionnée par la sociologie des réseaux [9]. L'analyse quantitative et structurale nous permettra en effet d'observer des régularités dans la forme que prennent ces relations.

Quant à la relation d'exploitation, elle relèverait d'une organisation délibérée par les auteurs – au sens juridique – des relations de dépendance d'une part et d'isolement d'autre part, des

8. On rappellera que le droit pénal assortit du prononcé d'une peine l'accomplissement de certains actes, limitativement énumérés dans le Code pénal, au motif qu'ils heurtent les valeurs considérées comme fondamentales dans une société.

9. La sociologie des réseaux consiste "à prendre pour objet d'étude non pas les attributs (âge, sexe, professions, etc.) ou les actions des individus, mais les relations entre les individus et des régularités qu'elles présentent, pour les décrire, rendre compte de leur formation et de leur transformation, analyser leurs effets sur les comportements individuels" (Mercklé, 2004).

personnes exploitées : relations non seulement avec les membres du groupe criminel, mais aussi avec l'entourage du pays d'origine (famille, relations interpersonnelles) et du pays de destination (autres personnes exploitées, société civile, administration, etc.). Ce deuxième pan de l'hypothèse sera questionné non seulement par la sociologie des réseaux et un traitement quantitatif des données, mais également par une analyse qualitative de celles-ci.

Les données étudiées permettent d'observer les relations sociales au sein d'un groupe criminel. Elles permettent de décrire d'une part, la relation d'exploitation *stricto sensu*, soit la relation qui unit celui qui exploite à celui qui est exploité et d'autre part, l'organisation des relations sociales qui environnent la relation pénalement répréhensible et renvoie alors à la situation d'exploitation. Au delà, les données founissent de nombreuses informations relatives à la description des acteurs, leurs rôles sociaux, leurs logiques d'action et trajectoires au sein de la pratique criminelle. Ce questionnement s'inscrit dans un contexte scientifique portant sur l'analyse des réseaux criminels qu'il importe de présenter.

1.2 Le contexte scientifique

La rareté des recherches empiriques portant sur les organisations criminelles se livrant à cette criminalité transfrontalière doit avant tout être soulignée. Cette rareté s'explique par la clandestinité des pratiques étudiées, la difficulté d'accès aux acteurs impliqués et la complexité de ces groupes (zones géographiques étendues, nombre d'auteurs impliqués, sophistication des modes de transfert de fonds, diversité des modes de communication, etc.) (Weitzer, 2014). Si des recherches importantes sur la quantification du nombre de victimes de traite des êtres humains en Europe ont été réalisées récemment (van Dijk *et al.*, 2014), le présent projet se situe davantage dans la lignée de recherches fondées sur un examen des dossiers actifs de la police.

Sur la base de ce type de données, on retiendra : une étude sur le lien entre grand banditisme français et les économies souterraines liées au trafic de drogue (Colombié *et al.*, 2001), une étude du lien entre prostitution et crime organisé à partir d'une photographie de formes de prostitutions identifiées à Genève (Sardi et Froidevaux, 2003), une recherche identifiant les liens entre trafic de drogue et réseaux d'immigration aux Pays-Bas, plaque tournante du trafic international de stupéfiants (Fijnaut *et al.*, 1998).

Néanmoins, ces différents travaux déjà anciens ne proposent qu'une approche parcellaire des questions abordées. Surtout, ils ne permettent pas d'avancer sur notre question de recherche liée à la définition de l'exploitation.

Dans ce domaine, les études portant sur la compréhension du mode opératoire des organisations criminelles se livrant à des faits de traite des êtres humains présentent pour intérêt de permettre d'identifier et de catégoriser les acteurs impliqués dans le processus d'exploitation. Les premiers travaux sont sans doute ceux de Salt et Stein (1997) qui ont mis en évidence trois phases dans le processus criminel : le recrutement, la phase migratoire et la phase d'immersion dans le pays de destination. Cette piste a ultérieurement été développée par Salt (2000). Ces auteurs ont présumé l'existence d'un personnage central supervisant le processus global. Par ailleurs, Aronowitz (2001) a établi une typologie d'organisations criminelles en fonction de leur taille. Selon elle, les organisations transnationales apparaîtraient en bout de chaîne. Ces organisations agissant sur plusieurs pays interviendraient du recrutement à l'exploitation en prenant en charge le transport, la fourniture des faux documents, la corruption des agents des douanes, etc. Le recours à la sociologie des réseaux, en plein essor dans le domaine de l'étude de l'activité criminelle permet d'aller plus loin sur ces questions (Calderoni, 2014; Carrington, 2014). Les

études les plus conséquentes ont été réalisées pour étudier le trafic de drogue (Bright *et al.*, 2014; Morselli, 2010; Calderoni, 2012; Malm et Bichler, 2011).

Dans le domaine de la traite des êtres humains, les principaux auteurs s'étant livré à une telle analyse sont Campana (2016) et Mancuso (2014). Ainsi, Paolo Campana a remis en question, en s'appuyant sur des dossiers judiciaires nigérians se livrant à des faits de traite, une partie des conclusions des travaux antérieurs, puisqu'il a montré une séparation entre les activités de transport et d'exploitation. Quant à Marina Mancuso, elle a montré l'existence de deux catégories de "Madams" - nom donné aux proxénètes nigérianes -. Alors que les unes ont une position hiérarchique importante, une forte centralité, et sont en mesure de surveiller toutes les phases du processus criminel d'exploitation, les secondes ont beaucoup moins de relations avec les autres membres de l'organisation criminelle et leur sphère d'influence est principalement limitée à l'exploitation *stricto sensu*. Leur position hiérarchique – structurale – est de ce fait plus faible.

Il importe enfin de souligner l'apport des travaux de Rossy par rapport à notre questionnement, travaux portant sur les techniques de cartographie et de visualisation des réseaux criminels (Rossy, 2011, 2016). Il démontre en effet les avancées, mais également les limites que peuvent revêtir ces techniques dans l'analyse criminelle. Dans la mesure où notre recherche comprend une dimension destinée à permettre la visualisation du réseau criminel, il est essentiel de repérer les implications du recours à de tels procédés.

L'étude réalisée devrait permettre d'approfondir l'ensemble de ces questions. Vérifier l'hypothèse d'une organisation maîtrisée des relations sociales au sein du groupe criminel renvoie à l'identification du rôle et du statut social de chaque individu apparaissant dans la procédure pénale ainsi que les liens entre eux. Le cumul ou le cloisonnement des tâches imputées aux acteurs devrait être mis en évidence avec précision. Plus encore, la démarche adoptée ici nous permettra d'aller au-delà de l'ensemble de ces travaux sur quatre points :

- L'élaboration pluridisciplinaire d'une méthodologie incarnée dans une plateforme spécifique rend possible la saisie et le traitement d'une quantité importante de données de nature qualitative, quantitative, structurale, géographique et temporelle. De cette manière, les processus et logiques d'actions y sont inscrits.
- Cette méthodologie vient également en appui d'une part à l'analyse exploratoire (Tukey, 1977; Thomas et Cook, 2006), débutant dès la collecte des données, et d'autre part à l'analyse confirmatoire (Brown, 2015) du réseau et de la pratique criminelle permettant ainsi de pallier techniquement et cognitivement la complexité et la fragilité du matériau de la recherche
- Elle permet de travailler sur l'ensemble des acteurs impliqués dans le réseau et non les seuls auteurs au sens juridique, c'est-à-dire, ceux dont les faits sont pénalement qualifiables. L'approche est donc globale et le statut de chaque acteur interrogé. De fait, elle sert la compréhension des interactions sociales mais aussi institutionnelles (services étatiques ou ONG) jouant de fait un rôle dans l'activité criminelle.
- Enfin, le traitement comparatif de plusieurs dossiers judiciaires est rendu possible par cette plateforme permettant à terme, d'observer les régularités et les facteurs déterminants des modes opératoires (taille, origine géographique, type d'activité, degré de contrainte, etc.)

La mise en oeuvre d'une telle approche a été rendue possible par l'association de trois disciplines au coeur de cette recherche.

La dimension pluridisciplinaire du projet. Les éléments qui précèdent mettent en évidence la dimension pluridisciplinaire de la recherche qui associe étroitement dans les éléments constitutifs de la problématique, juristes et sociologues. Néanmoins, l'ampleur des données à traiter a nécessité le recours à des chercheurs en informatique en vue de la création d'un outil spécifique facilitant l'extraction (et la saisie) mais également l'analyse et la visualisation d'une quantité importante de données. L'association de ces trois disciplines a conduit à l'élaboration d'un langage commun, à l'identification des difficultés techniques, des biais d'analyse rencontrés et des moyens de les surmonter.

En outre, le caractère pluri-disciplinaire de la gestion des données et de l'étude a imposé d'adopter une démarche incrémentale pour asseoir le processus de collecte et d'élaboration d'un modèle de données informatique suffisamment robuste pour accompagner l'étude. S'est installé un jeu de va-et-vient permanent entre le recueil analytique des données et le contenant informatique les stockant. Il faut voir dans ce défi le processus même qui permet d'accoucher simultanément d'un modèle pertinent pour le juriste (au sens de la problématisation de la recherche) et d'un modèle de données informatique cohérent (vu comme un contenant rigoureux).

Si ce travail revêt une portée théorique mais également potentiellement, opérationnelle, la présente contribution se concentrera sur la dimension méthodologique. Partant, il importe de présenter les moyens mis en œuvre, entendus ici comme une mise à disposition de la "science des réseaux" au service de la science juridique (Section II), et d'exposer enfin les premiers observations analytiques(Section III) et discussions méthodologiques (Section IV).

II LA SCIENCE DES RÉSEAUX AU SERVICE DES SCIENCES JURIDIQUES

La dynamique globale de cette étude est de permettre la vérification d'une hypothèse de recherche à forte implication juridique, en s'appuyant sur une méthodologie d'une part sociologique (recueil et analyse)(Section 2.1) et d'autre part apportée par la science informatique (plateforme de saisie, modélisation et visualisation) (Section 2.2)

2.1 Mobiliser la sociologie et l'analyse de réseaux sociaux

La sociologie a été mobilisée en premier lieu dans sa capacité à mettre en œuvre une méthodologie de recherche permettant d'associer une problématique de recherche à un travail de terrain empirique. Qu'il s'agisse d'une approche hypothético-déductive ou inductive, il est nécessaire de définir les concepts-clés et les indicateurs pertinents associés au matériau et à la question de recherche. Cette procédure d'opérationnalisation des concepts (Lazarsfeld, 1965) relève de l'élaboration d'une méthodologie de recherche d'autant plus importante ici que les données sont de nature judiciaire et qu'elles relèvent d'une analyse dite secondaire (Section 2.1.1). Or, développer une méthodologie de recherche suppose la reformulation de la problématique originairement juridique. Aussi, en apportant de nouveaux outils et points de vue, la sociologie a enrichi la recherche par de nouvelles hypothèses et questionnements, soulevés notamment par une perspective structurale des données proposée par la sociologie des réseaux (Section 2.1.2).

2.1.1 Opérationnaliser sur la base de données d'investigation

Les données étudiées relèvent d'une analyse dite secondaire [10] : elles ont été produites et collectées à des fins autres que la recherche scientifique, à savoir en l'espèce, à des fins d'investigations policières.

10. L'analyse secondaire est définie comme étant la ré-exploitation de données d'enquêtes dont les résultats prolongent et se distinguent de l'analyse originaire, issue du recueil des données (Dale, 1993).

Ce point est à l'origine d'un certain nombre de biais dans le cadre de l'utilisation à laquelle la recherche les destine. De ce fait, il a été nécessaire d'une part de prendre en compte leur nature spécifique et les biais associés et d'autre part de procéder à des choix méthodologiques rigoureux afin qu'elles deviennent des données de recherche à part entière, capturant les divers éléments – entités et événements du réseau criminel – à même de confirmer ou non l'hypothèse de recherche développée.

L'identification du matériau de recherche. Un dossier judiciaire rassemble l'ensemble des pièces constituées au cours des phases de l'enquête puis de l'instruction pénale. Il comprend principalement les pièces suivantes :

— Procès verbaux d'auditions / interrogatoires (services police /magistrats) accueil [11] ;
— Retranscription d'écoutes téléphoniques accueil [12] ;
— Pièces obtenues sur réquisitions accueil [13] ;
— Documents saisis accueil [14] ;
— Compte rendu de transport / surveillance visuelle ;
— Diverses pièces de procédure ;
— Procès-verbaux de synthèse [15].

Recueillies à des fins d'investigation judiciaire, les données ne présentent ni l'homogénéité ni l'exhaustivité que l'on pourrait attendre dans le cadre d'une démarche scientifique. Les enquêteurs cherchent à prouver que des actes pénalement répréhensibles ont été commis. Leurs investigations se focalisent donc sur les seuls éléments susceptibles de se rattacher aux faits poursuivis. C'est ce qui explique que lors des écoutes téléphoniques, les traducteurs tendent à filtrer les échanges retranscrits pour ne retenir que les informations directement utiles à l'enquête. De même, certains choix lors de l'enquête se justifient par des critères économiques ou managériaux : caractère onéreux et chronophage des écoutes téléphoniques ; travail à flux tendu limitant la possibilité de développer tel ou tel aspect du dossier, etc.

Ainsi, on ne connaîtra jamais que la nationalité de certains individus sans jamais déterminer leur lieu de naissance ; pour d'autres on aura une adresse complète du lieu de résidence (rue, ville, code postal) quand on ne pourra renseigner que le nom d'un pays pour d'autres encore. De la même manière, la date d'une interaction entre criminels sera parfois renseignée au jour près $(jj/mm/aaaa)$; parfois on indiquera le mois et l'année ; et dans une autre hypothèse, on ne pourra qu'affirmer qu'elle a eu lieu avant l'année $aaaa$, ou entre les mois mm et mm' de l'année $aaaa$.

S'agissant du contenu même de ces données, par définition, les informations issues dans les pièces de procédure doivent être décryptées : les protagonistes utilisent de nombreux alias pour ne pas être identifiés, ils mentent aux enquêteurs et utilisent entre eux, un langage codé. Seule une lecture qualitative armée d'une connaissance thématique préalable permet de comprendre le sens des données recueillies.

11. Procès verbaux sur lesquels sont retranscrits les auditions des victimes, témoins ou personnes mises en causes par les enquêteurs ou magistrats.
12. Document dans lequel sont retranscrites les conversations liées à un numéro de téléphone placé sur écoute.
13. Opérateurs de téléphonie, banques, coopération pénale internationale, documents d'état civil ou liés à la situation administrative, etc.
14. Extraits actes de naissance, passeports, relevés de transferts de fonds, documents manuscrits, etc.
15. documents rédigés par les enquêteurs pour faire état de l'avancement de leurs investigations.

Cette recherche revêt donc également un enjeu méthodologique considérable. Le caractère polymorphe de certains attributs décrivant les lieux ou le temps, le fait que les données sont parfois incomplètes, incertaines ou incohérentes – comme c'est souvent le cas des réseaux criminels (Xu et Chen, 2005) – est une des majeures difficultés à résoudre. En outre, ces éléments démontrent qu'une extraction automatique n'était pas envisageable.

L'analyse secondaire de données. Comme énoncé précédemment, nous n'avons pas élaboré un outil de recueil des données en adéquation avec notre problématique de recherche (guide d'entretien, questionnaire, générateur de noms, etc.) ; nous avons recueilli les données nécessaires à l'analyse dans des documents déjà constitués et non actualisables. Ce projet renvoie donc à la nécessité de toujours penser la donnée saisie dans ses potentialités analytiques, de répondre à des hypothèses y compris de manière inductive.

De par leur nature opérationnelle, les données sont très hétérogènes : contenus de discussion personnelle, interrogatoires, documents administratifs, traces de flux géographiques et financiers, etc. Aussi, elles ont été saisies de manière quasi exhaustive et organisées *a posteriori* afin de pouvoir fournir des éléments de réponse à la question initiale posée par le droit. Cette diversité rend complet et complexe le matériau de recherche ainsi obtenu. Une analyse qualitative, statistique et structurale est ainsi réalisable et permet une démarche analytique globale de la pratique criminelle. Au-delà de l'analyse du discours des acteurs interrogés, nous avons en effet des données relationnelles (A est en relation avec B), quantitatives (caractéristiques des individus et des relations) et qualitatives (nature des relations, contextes et environnements de création et de développement de celles-ci).

Compte tenu de la nature des données précédemment énoncée, on observe un nombre important de données manquantes liées aux attributs des acteurs (critères socio-démographiques), le manque de fiabilité des données récoltées et le focus sur un nombre réduit d'acteurs.

Néanmoins, deux éléments limiteront à terme la portée de ces biais. D'une part, la multiplication des dossiers, et donc du volume de données, permettra de réduire le poids des données manquantes. D'autre part, il s'agit d'une analyse préliminaire à l'issue de laquelle sera déterminée la pertinence des variables et leur sélection finale dans l'élaboration du traitement analytique (statistique et analyse de réseau). La quantité de données manquantes pourra alors être un indicateur de sélection de la variable.

2.1.2 Le recours à la sociologie des réseaux

Sociologiquement, la question posée est la suivante : qu'est ce qui tient socialement, structurellement et individuellement, les victimes dans ce système social ? Autrement dit, qu'est-ce qui caractérise ce système ? Comment fonctionne-t-il ? Qui sont les individus qui le composent et quelle est leur importance, leur position dans ce système ? Au final, il s'agit de mettre à jour les différents modes de fonctionnement et d'organisation sociale des réseaux criminels et d'en définir une typologie. Pour ce faire, le groupe criminel est considéré comme un "réseau criminel" dont il faut définir les frontières et décrire les actions fonctionnelles qui le constituent.

Analyse d'un réseau criminel. Juridiquement, on entend par "groupe criminel organisé", "un groupe structuré de trois personnes ou plus existant depuis un certain temps et agissant de concert dans le but de commettre une ou plusieurs infractions graves (...) pour en tirer, directement ou indirectement, un avantage financier ou un autre avantage matériel" [16]. En droit interne,

16. Article 2 a) de la Convention des Nations unies de lutte contre la criminalité transnationale organisée", adoptée par la Résolution 55/25 de l'Assemblée générale du 15 novembre 2000.

la bande organisée visée par l'article 132-71 du Code pénal suppose la préméditation d'une part, et que l'on puisse démontrer "une organisation structurée entre ses membres" d'autre part [17]. Autrement dit, celle-ci implique "un agencement des membres et une coordination tournée vers la commission d'une infraction déterminée".

En sociologie, le terme d' "organisation", en tant qu'objet social, renvoie à un "ensemble humain ordonné et hiérarchisé en vue d'assurer la coopération et la coordination de leurs membres pour des buts donnés" (Besnard *et al.*, 1999). Les buts, mécanismes de contrainte et modes de légitimation de l'autorité diffèrent selon les organisations. Néanmoins, elles impliquent toutes que leurs membres fassent preuve d'un "minimum de coopération indispensable à leur survie". Or le caractère construit – et donc non naturel – d'une organisation repose sur une triple limitation : les membres de celle-ci ne sont jamais complètement dépendants les uns des autres et ont une marge de liberté qu'ils cherchent à défendre (Crozier et Friedberg, 1992) ; la rationalité des comportements de tous les acteurs repose sur des visions locales et partielles (March et Simon, 1999) sans qu'aucune rationalité supérieure et englobante ne coïncide ; la faible capacité d'intégration de l'organisation est en concurrence avec les objectifs et intérêts de chacun des membres (Silverman, 1973).

Notre objet d'étude ne peut entrer dans ces modalités analytiques. Dans le contexte des réseaux criminels liés à la prostitution, nous posons les hypothèses opposées d'une interdépendance forte des acteurs ; de l'absence de marge de liberté ; d'objectifs et rationalités communs et englobants. Aussi, la nature du matériau et le questionnement développé ont inévitablement orienté la recherche vers la sociologie relationnelle et plus précisément la méthodologie portée par la sociologie des réseaux sociaux qui consiste à prendre la relation sociale comme point de départ pour étudier les phénomènes sociaux. Elle permet de ne pas aborder les individus et la société – ici criminelle – comme deux entités distinctes et antagonistes.

Cette approche permet de mettre en évidence l'existence d'une organisation maîtrisée de l'ensemble des relations entre les individus prenant part à une activité criminelle quel que soit leur rôle et position au sein de celle-ci. Cette démarche implique de penser les informations recueillies par les enquêteurs en données exploitables dans le cadre d'une analyse de réseaux.

Le groupe criminel est ainsi considéré du point de vue de l'objet "réseau" ici défini comme "un ensemble d'unités sociales et des relations que ces unités sociales entretiennent les unes avec les autres, directement, ou indirectement à travers des chaînes et des chemins relationnels de longueurs variables" (Mercklé, 2004), et permettant la circulation de ressources (Grossetti, 2009).

Ainsi, le projet repose sur l'analyse de réseaux d'acteurs identifiés dans un dossier judiciaire ayant donné lieu à un jugement définitif sanctionnant des faits de traite des êtres humains. On entend par acteur, l'ensemble des individus identifiés, interrogés et/ou cités dans la procédure, y compris, de ce fait, les personnes morales qu'il s'agisse d'acteurs associatifs ou d'agences immobilières par exemple. A ce stade, il s'agit donc d'identifier l'ensemble des acteurs qu'ils jouent ou non un rôle direct ou indirect dans le processus d'exploitation. C'est notamment sur ce point que le travail entrepris va plus loin que les études de Campana (2016) et Mancuso (2014) qui ont limité leur analyse aux personnes dont l'activité relevait de la qualification pénale d'auteur ou de victime (soit respectivement 58 et 86 individus).

17. Porteron C., Note sous Crim. 8 juillet 2015, no 14-88.329, Actualité Juridique Pénal 2016, p. 141.

Dès lors, le réseau identifié ne sera pas un réseau criminel mais le réseau des acteurs identifiés dans une procédure judiciaire visant la qualification de traite des êtres humains. Cela n'exclut pas que les mêmes individus puissent être identifiés dans d'autres procédures judiciaires. Cela n'exclut pas davantage que des acteurs puissent être impliqués dans d'autres activités criminelles que l'exploitation (Campana, 2016) : migration illégale, stupéfiant, terrorisme, etc.

Un réseau d'actions fonctionnelles. La transformation des données judiciaires en données de recherche a donc impliqué d'extraire l'ensemble des acteurs interrogés ou cités lors des interrogatoires afin de formaliser et contextualiser les liens décrits entre ces acteurs. Concrètement, le sociologue va concentrer son regard sur les interactions et relations, il va générer des noms, des *ego* et des *alter*, leurs caractéristiques, et les descripteurs de la relation (lieu, date, etc.). Il faut cependant disposer des outils nécessaires pour mener à bien un recueil rigoureux et systématique de l'ensemble des données et rassembler les données sur les acteurs et leurs relations. Pour cela, il faut définir les acteurs concernés, les frontières du réseau et la nature des liens qui le compose, le tout, en gardant en tête notre problématique.

Pour rappel, nous avons défini qu'un acteur désigne toute personne qui a un rôle direct ou indirect, conscient ou inconscient, dans le processus d'exploitation, y compris le soutien associatif. Pour être pris en compte, un individu doit être suffisamment identifiable par son prénom, nom ou alias. Cette précision indique un niveau de connaissance minimum (Strauss, 1992) et porte déjà un certain nombre d'éléments caractérisant socialement, individuellement et biologiquement l'individu (Bourdieu, 1986). Lorsqu'une conversation ou une audition évoque simplement "quelqu'un", "un homme" ou "un ami",... ledit individu n'est pas retenu.

Un certain nombre d'indicateurs socio-démographiques permettent de caractériser chaque acteur : nom, prénom, sexe, alias, langue(s) parlée(s), numéro(s) de téléphone, date de naissance, statut familial, lieu de naissance, nationalité, profession, niveau de diplôme, situation familiale et fratrie, enfants etc. Nous avons également la possibilité de quantifier, par l'analyse, les localisations géographique, nombre de déplacements, nombre de"filles" travaillant pour lui, etc.

Le réseau, quant à lui, est circonscrit au dossier judiciaire : le réseau des acteurs identifiés dans une procédure judiciaire visant la qualification de traite des êtres humains. Plus encore, il s'agit de l'agrégation des réseaux personnels des individus interrogés plutôt qu'un réseau criminel à part entière : c'est la forte multiplexité [18] des relations qui permet de relier ces réseaux égocentrés entre eux.

Enfin, concernant le type de lien et suivant notre question de recherche, nous abordons le réseau sous tous ses aspects relationnels. La pratique criminelle repose sur des tâches, sur des actions qui la constituent. Elle est encastrée (Granovetter, 1985) au sein d'un réseau de relations personnelles. Nous avons donc focalisé notre regard sur les interactions fonctionnelles au cœur de l'activité criminelle – il s'agit d'un "Action-based Network" –. Autrement dit, les liens qui unissent les acteurs sont principalement basés sur les actions : "A fournit une place de trottoir à B" ; "B se prostitue pour C". Il y a autant de liens qu'il y a d'actions et d'interactions fonctionnelles entre les acteurs du réseau criminel. En outre, la mutiplexité des liens est ici importante et structure de la même manière le réseau et l'activité criminelle. Aussi, nous avons également pris en compte les liens familiaux, matrimoniaux etc. : "A est la sœur de D" ; "D est marié avec

18. Il s'agit d'un néologisme élaboré par Mitchell (1969) pour mettre en évidence le fait qu'une relation dyadique peut engendrer un conflit ou une complémentarité entre deux rôles et deux parties du réseau de l'individu. Autrement dit, une relation peut s'incrire dans plusieurs contextes relationnels et sociaux. Si votre frère est aussi votre collègue, la relation s'inscrit dans deux contextes : familial et professionnel.

C". Une relation entre deux acteurs peut en effet s'opérer simultanément dans ces différents cercles recouvrant ainsi différentes réalités sociales et criminelles. De cette manière et à des degrés différents, "A, B, C et D sont en liens".

Partant, nous avons défini sept types de liens permettant la formalisation du réseau criminel.

— On entend par "liens de réseau" toutes interactions/actions/contextes liés à l'exploitation et à la prostitution permettant de définir, dans le cadre de l'activité criminelle, un lien entre deux individus ou plus – il peut s'agir de chaînes de relations avec des intermédiaires. Ces relations liées à l'exploitation sont par exemple associées aux actions de type "fournit une place de trottoir / bénéficie d'une place de trottoir" ; "se prostitue pour / sponsorise [19]" ; "initie à la prostitution / est initiée" ; "recrute au pays d'origine / est recrutée", etc.
— Les "liens financiers" sont, quant à eux, des relations qui reposent sur un flux financier déclaré ou observé entre deux individus. Un échange financier a lieu en contrepartie d'une action liée à l'activité criminelle. Contrairement aux "liens de réseau", il y a autant de mentions du lien que de preuves de la transaction. Une même relation entre deux acteurs peut donc être identifiée à de nombreuses reprises, puisqu'à chaque fois, correspond un nouvel échange d'argent. La densité de lien observé ici ne relève donc pas du même niveau d'analyse.
— Les "liens de soutien" sont des relations liées aux activités de soutien. Cette catégorie correspond à une relation bimodale, puisqu'elle unit un individu à une personne morale (généralement une association venant en aide aux personnes migrantes ou prostituées). Il ne s'agit donc pas d'un lien interindividuel.
— Les "liens de sang" relèvent de liens de filiation prétendus ou avérés et soulèvent des enjeux culturels fortement impliqués dans leur désignation.
— Les "liens sexuels" sont des relations définies sur la base d'une activité sexuelle et/ou de couple ; il peut s'agir d'un lien avec le conjoint ou un client, par exemple.
— Les "liens juju" [20] sont un type de liens d'une part spécifiques aux réseaux d'exploitation nigérians et d'autre part, ne correspondant pas à des relations dyadiques mais à la présence de chacun des acteurs à une cérémonie "juju".

On peut avoir une relation de type "réseau" en même temps que "financier" avec un membre de sa famille présent lors de la même cérémonie du juju. On peut se prostituer pour sa cousine par exemple, lui rembourser sa dette régulièrement alors même qu'elle était le témoin lors de la cérémonie du juju. Seuls les "liens de connaissance" – dernier type de liens observé – sont exclusifs des liens précédemment cités : ce sont des relations définies par un lien de connaissance minimum renseigné par autrui ou par les protagonistes eux-mêmes sans qu'aucune action ou lien de filiation ne soit – encore – directement associé à ce lien. Ces derniers permettent ainsi de générer la structure la plus exhaustive possible.

La plupart des liens décrivent l'action des deux protagonistes de la relation : celle de l'*ego* qui est en position active et celle de l'*alter*, qui est plus fréquemment – mais non systématiquement

19. On désigne par "sponsor" toute personne faisant l'avance des frais liés à la migration et à l'activité prostitutionnelle.

20. "Les pratiques juju sont des rites de magie noire au cours desquels des vêtements intimes sont enlevés, des tissus et des fragments du corps et des fluides corporels (par exemple des poils pubiens, des cheveux, des ongles et du sang menstruel) sont prélevés sur les femmes et placés dans un lieu saint. [...]. Ces rites sont d'une grande importance pour les victimes car elles sont profondément convaincues que le mauvais sort s'abattra sur elles et sur leurs familles si elles ne remboursent pas leurs dettes." (Aghatise, 2005).

– en position passive. Pour chaque action, la date – lorsqu'elle est connue – est renseignée, permettant à terme une vision dynamique et diachronique du réseau.

Ces éléments suscitent d'importantes difficultés, puisqu'on l'a vu, la précision des informations est très variable. Malgré ces difficultés, les données recueillies permettent d'ores et déjà des avancées considérables quant à l'assise scientifique des éléments de description des réseaux criminels. Et ceci ne saurait être possible sans l'apport de la science informatique.

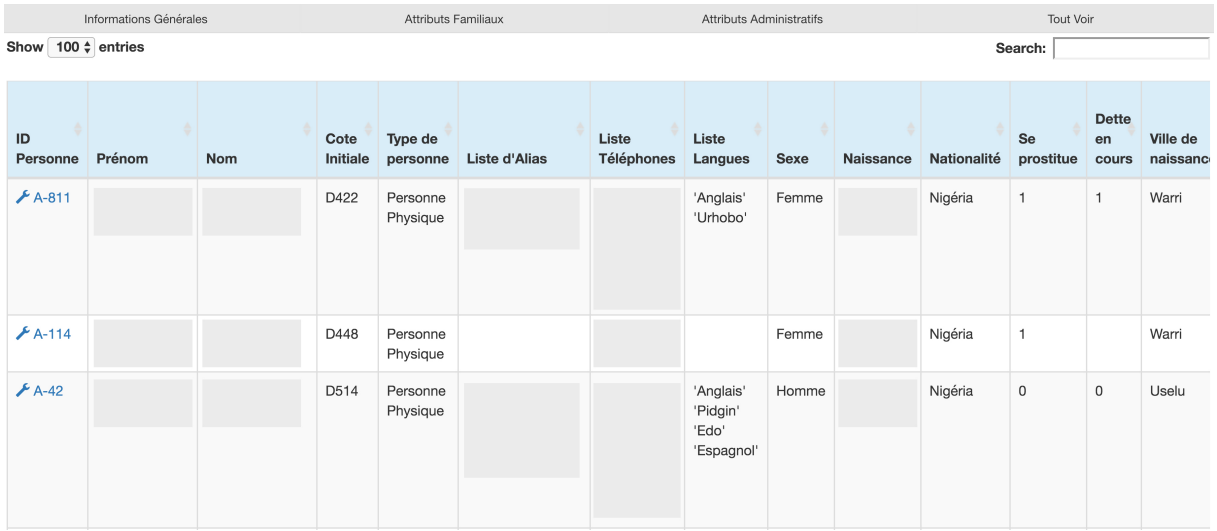
2.2 Mobiliser la science informatique

L'analyse des données appelle de manière évidente le soutien de l'informatique. En amont, l'informatique apporte à la méthodologie entreprise des exigences en termes de stockage des données (Section 2.2.1). A terme, le travail de saisie anticipe la phase de l'analyse. Il autorise en effet une validation en temps réel : les données saisies sont comparées aux données déjà présentes dans la base (et visualisées au moment de la saisie), et cette navigation répétée amène les experts vers les prémisses du modèle du réseau (Section 2.2.2).

2.2.1 Soutenir la collecte

Parce que la démarche des chercheurs est par nature exploratoire, il importe d'extraire toutes informations ou données susceptibles d'être utiles. L'extraction est un processus chronophage et coûteux et on ne peut envisager de revenir sur les documents d'origine pour compléter la collecte qu'au prix d'un effort considérable. Les acteurs du réseau sont équipés d'attributs – critères socio-démographiques. Un acteur donne lieu à un peu moins d'une centaine d'attributs à renseigner, décrivant son identité, sa localisation (incluant ses éventuels déplacements) et sa situation familiale ou administrative. Les relations entre les acteurs sont également caractérisées par des attributs divers décrivant des éléments de contexte, de localisation et/ou de datation.

D'abord effectuée à l'aide de tableurs classiques, il est vite apparu essentiel de concevoir et réaliser un outil d'aide à la saisie. Le schéma relationnel, tabulaire, a donc servi de base à la création d'écrans de saisie (Figure 1) dont l'un des rôles était de restituer les informations déjà recueillies sur les personnes – ces informations étant assemblées à mesure de l'examen des dossiers.


Informations Générales | Attributs Familiaux | Attributs Administratifs | Tout Voir

Show 100 entries — Search:

ID Personne	Prénom	Nom	Cote Initiale	Type de personne	Liste d'Alias	Liste Téléphones	Liste Langues	Sexe	Naissance	Nationalité	Se prostitue	Dette en cours	Ville de naissanc
A-811			D422	Personne Physique			'Anglais' 'Urhobo'	Femme		Nigéria	1	1	Warri
A-114			D448	Personne Physique				Femme		Nigéria	1		Warri
A-42			D514	Personne Physique			'Anglais' 'Pidgin' 'Edo' 'Espagnol'	Homme		Nigéria	0	0	Uselu

FIGURE 1 – Vue partielle d'un écran de saisie reprenant le schéma relationnel de données (sur les personnes).

La complexité des données décrivant le réseau tient au caractère multi-attribué des entités qui le composent et à la variété des types d'interactions qui y prennent place. Ces données polymorphes posent un défi en termes de stockage et de traitement. Par exemple, la souplesse attendue au niveau des dates exige de les stocker sous forme de chaînes de caractères amputant ainsi le processus de stockage des mécanismes propres aux données typées. Formuler un modèle de données rigoureux incarné dans une base de données permet d'assurer l'intégrité et la cohérence des données : la gestion d'identifiants uniques pour les personnes et les relations évite la saisie de doublons ; les variations orthographiques peuvent être évitées dès lors que la saisie s'appuie sur les informations déjà présentes dans la base.

La taille de la base est maîtrisée en évitant de dupliquer des entités secondaires renseignant les attributs des personnes ou des liens (côtes des documents décrivant l'origine de la donnée, nationalité des personnes, nom des pays, villes, numéros de téléphone, etc.).

Très vite, il est apparu essentiel que l'interface de saisie puisse également devenir un outil de consultation des données. En effet, une nouvelle information doit souvent être confrontée à une information déjà saisie. Par exemple, un même alias de nom peut être utilisé par deux personnes distinctes ; à l'inverse, on réalisera qu'une personne dont on ne connaissait que l'alias est celle qui avait été renseignée sous son nom propre connu dans d'autres documents.

Ces simples exemples illustrent bien les exigences de malléabilité imposées au schéma de données et à l'interface de saisie. Il a fallu fusionner les informations sur deux personnes au moment où l'on découvre qu'il s'agit d'un seul et même individu. De la même manière, les modalités des variables observées ont souvent eu à être "externalisées" et stockées dans une table d'association distincte pour leur donner une flexibilité maximale, avant d'être factorisées et simplifiées au moment de l'analyse.

2.2.2 Le modèle relationnel

La complexité du schéma tient aussi en partie au caractère inductif de la démarche de l'analyse. Les tables d'association ont été ainsi utilisées, et souvent remaniées, un peu à l'image de notes "post-it" que l'on accolerait au fil de la lecture des dossiers, avant d'en faire une synthèse après avoir pris le recul nécessaire.

Motivé par les approches classiques de la sociologie, le choix a été fait de développer un modèle de données reprenant une vision tabulaire entités/attributs (voir Figure 2). Nous avons ainsi établi un modèle de données relationnel, au sens informatique.

Dans la Figure 2, les relations sur la gauche stockent les informations relatives aux relations. Les relations en haut à droite renseignent les personnes. Au centre, se trouvent les attributs partagés entre les relations et les personnes, renseignant les attributs qui ont trait aux localisations géographiques (adresse, nationalité, passeport, etc.), aux professions, numéros de téléphones et alias, etc. Le schéma s'étend sur un peu moins de 50 relations (tables) :

— Huit relations (tables) rassemblent plus de attributs sur les personnes (identité, situation familiale, administrative, langue/s parlée/s, alias de nom, téléphone/s, profession/s, rôle/s dans le réseau).
— Huit relations (déclinant les liens par types) rassemblent un peu moins d'une centaine d'attributs sur les liens entre personnes.
— Six relations (tables) rassemblent des données décrivant des informations relatives à la localisation et aux déplacements.

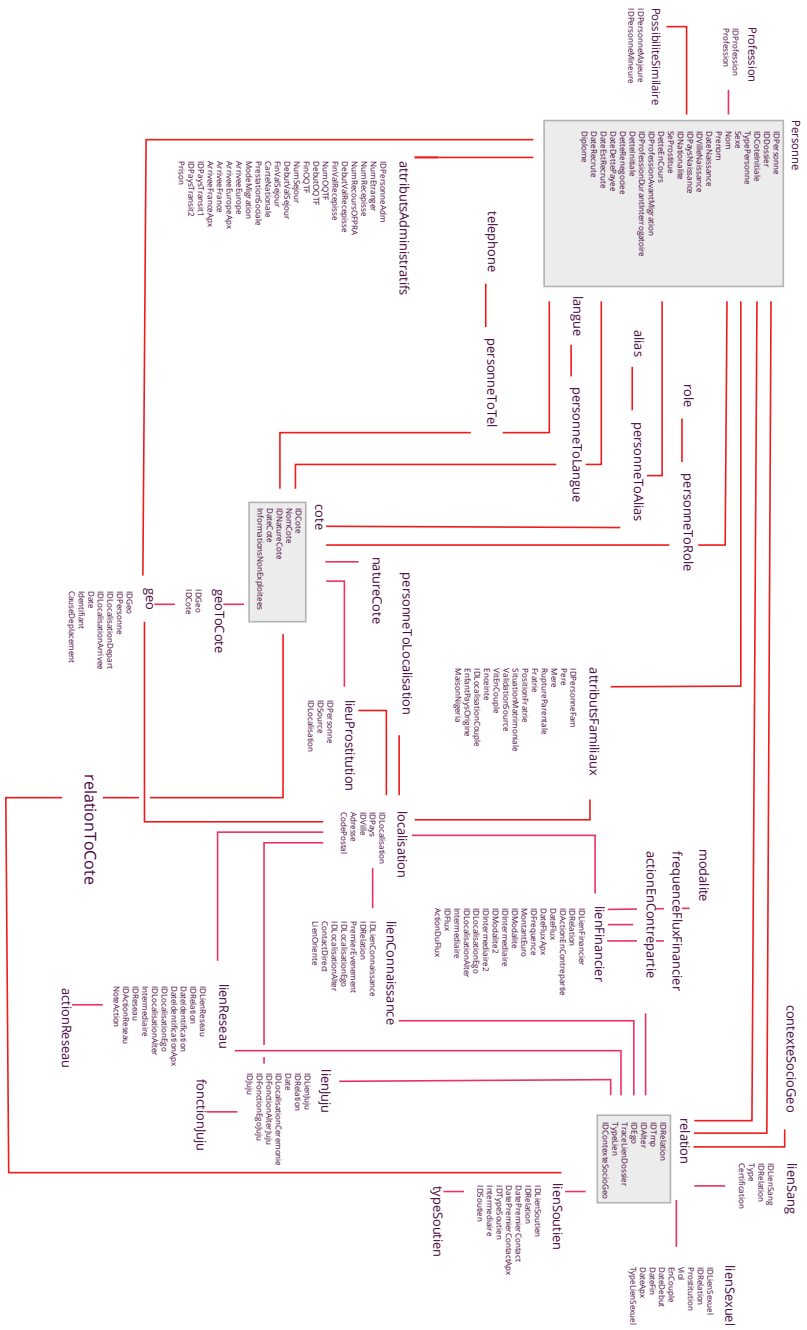


FIGURE 2 – Vue d'ensemble du schéma relationnel de données (simplifié).

— Sept relations additionnelles viennent enrichir les informations qualifiant les liens entre personnes (action posée, modalité et fréquence de paiement/s, type de soutien apporté).
— Le schéma contient plus d'une dizaine d'associations (mettant en relation des éléments décrits dans des tables distinctes) équipés d'attributs propres.

2.2.3 Accompagner l'analyse

Les données collectées et stockées ne sont pas filtrées ou "normalisées". Les données sont stockées à l'état "brut" pour donner à l'analyste la possibilité d'en dériver un sous-ensemble ou un sous-réseau particulier "à la demande". Un sous-réseau pourra ainsi être induit de critères géographiques ou temporels, de forme d'exploitation, d'interaction avec d'autres types d'ac-

tivités illicites, etc. On pourra ainsi isoler et étudier des sous-réseaux, en adaptant les critères d'analyse de la position structurale des acteurs (mesures de centralité) et de leur forme (densité, cohésion, connexité, transitivité, etc.).

La valeur ajoutée d'un modèle de données tient à sa capacité à épouser le plus fidèlement possible la nature des données à stocker (Section 2.2). L'outil informatique venant en appui à la tâche d'extraction des informations doit répondre à diverses exigences en termes de maniabilité et de réactivité. Il doit aussi accompagner l'intention de l'analyste, c'est-à-dire se soumettre de la manière la plus souple possible au traitement envisagé, et ultimement être un matériau naturel, à en faire oublier le support informatique (Munzner, 2009; Meyer *et al.*, 2012). La nature des activités du réseau criminel, l'ensemble des données récoltées, et les questions posées nous amènent dans le champ de la science des réseaux – "Network Science" telle que définie par Barabási (2011) ou Newman *et al.* (2006) par exemple. Le réseau criminel considéré n'est pas "un" réseau mais résulte en réalité de l'assemblage de plusieurs réseaux capturant des liens de natures diverses (section 2.1). La notion de réseaux multi-couches (Osusky, 2007; Kivelä *et al.*, 2014) est ici des plus pertinentes puisque les liens entre personnes sont typés. Elle épouse au mieux les caractéristiques d'origine des données et apporte toute la souplesse nécessaire à leur exploration et à leur exploitation analytique. Chaque type de lien induit un sous-graphe qui peut être étudié en soi, ou qui peut être composé avec les autres couches à des fins d'analyse. Ainsi, on pourra comparer la structure du réseau d'interaction à celle du réseau de filiation pour tenter de cerner la place ou le rôle des familles dans l'ensemble des activités du réseau criminel.

Les réseaux multi-couches constituent donc une abstraction utile. Ils constituent un artifice de visualisation pertinent pour développer une vision d'ensemble des données collectées, tout comme une représentation graphique intuitive pour formuler des requêtes lors de l'analyse. Néanmoins, il existe à ce jour encore peu de métaphores graphiques présentant les "couches" des réseaux. Les travaux de de Domenico *et al.* (2015) proposent de présenter ces réseaux en empilant les couches dans une visualisation 2D1/2. Or, les problèmes d'occlusion connus de ces approches suggèrent d'explorer d'autres pistes. Nous avons opté dans les premières phases de travail présentées ici des vues simultanées des couches, conjuguées à la possibilité de fusionner certaines couches en un seul sous-réseau "à la demande". Notre approche mise sur la possibilité donnée à l'utilisateur d'agir sur les différentes vues du réseau pour en questionner la structure ou accéder aux données sous-jacentes.

III HYPOTHÈSES ET PREMIÈRES OBSERVATIONS ANALYTIQUES

En termes juridiques, ce projet est porté, on l'a vu, par l'ambition de mettre à jour ce qui caractérise la notion d'exploitation : la relation proprement dite et la situation dans laquelle elle s'exerce. L'hypothèse formulée est ainsi qu'au-delà de paramètres contingents comme le type d'exploitation, la taille du réseau, l'origine géographique des individus qui le composent, la relation d'exploitation relèverait d'une organisation maîtrisée de l'ensemble des relations entre les individus prenant part à une activité criminelle quel que soit leur rôle et position au sein de celle-ci. Plus précisément, la relation d'exploitation relèverait d'une organisation délibérée par les auteurs – au sens juridique – des relations de dépendance d'une part et d'isolement d'autre part, des personnes exploitées Cette hypothèse de recherche s'étaye en de nombreuses autres hypothèses. Certains éléments de pré-analyse permettent d'ores et déjà de légitimer leur formulation voire d'apporter quelques réponses. Néanmoins, les premiers résultats inscrits ici ne sont en aucun cas généralisables ; il s'agit d'un travail d'analyse en cours permettant davantage de décrire et de circonscrire la recherche. Ici, nous nous appuyons sur le recueil et une première analyse descriptive des données d'un dossier judiciaire. C'est sur la base de ce dossier que nous

proposons le modèle d'analyse (Section 3.1) et les premières cartographies du réseau (Section 3.2).

3.1 Une analyse mixte et multiscalaire

La nature du matériau recueilli ainsi que l'approche théorique et méthodologique entreprise – soit le recours à la sociologie des réseaux – nous permettent de mettre à jour les rôles des acteurs du réseau et leur position structurale au sein de ce dernier, les types et les contenus des relations et d'interactions entre eux et enfin la forme de la structure du système relationnel. Ainsi, nous pourrons penser notre objet d'étude selon plusieurs échelles et à partir d'une analyse qualitative, quantitative et structurale des données. Les hypothèses formulées rejoignent ainsi les trois niveaux d'analyse (micro, meso, macro soit l'acteur (Section 3.1.1), la dyade (Section 3.1.2), le réseau (Section 3.1.3)) inhérents à la problématique et intègrent des analyses complémentaires.

3.1.1 L'analyse de l'acteur

Le premier groupe d'hypothèses procède de l'analyse des individus qui composent le réseau criminel. Indépendamment de toute référence à la qualité juridique d' "auteur" d'une infraction pénale, c'est-à-dire ici, par référence aux faits visés à l'article 225-4-1 du Code pénal, se pose la question de la place de l'individu au sein de ce système. Quel rôle joue-t-il dans l'activité criminelle d'une part, mais également dans le développement et le maintien du système d'autre part ? Qu'est-ce qui fait que l'individu va jouer un rôle ? Ce niveau d'analyse nous permet d'appréhender les caractéristiques socio-démographiques de chaque acteur [21], son rôle social – entendu comme sa fonction sociale au sein de l'activité criminelle –, son expérience et ses logiques d'action, sa position structurale au sein du réseau – entendu comme le rôle structural, soit la mesure de sa centralité–. Pour les acteurs interrogés dans le cadre de la procédure judiciaire, il est également possible de réaliser une analyse de leur réseau personnel. L'ensemble de ces éléments est appréhendé conjointement par l'analyse du contenu qualitatif, l'analyse statistique des données sociales et relationnelles, et l'analyse structurale.

Au sein du premier dossier judicaire traité, 318 acteurs ont été identifiés comme cités ou interrogés : 9 personnes morales (association, foyer, agence immobilière, etc.) et 309 personnes physiques (toute personne qui a un rôle direct et/ou indirect dans le processus d'exploitation, mais également enfants, client, bailleur, . . .). Les deux-tiers des individus sont des femmes. La moyenne d'âge est de 29 ans.

L'action est au fondement du réseau et *a fortiori* de l'objet d'étude : c'est à partir de son analyse que nous pouvons formaliser et caractériser les liens entre les acteurs, les fonctions et rôles sociaux que ces derniers endossent et le réseau social dans son ensemble. L'action criminelle est donc ici relationnelle ; elle prend corps dans une interaction fonctionnelle qui dans un système d'attentes réciproques l'oriente et associe ainsi l'individu au rôle attendu. Nous avons observé trente-cinq actions qualifiées de "réseau" au sens où elles sont principalement constitutives de l'activité criminelle (commande/recrute au pays d'origine ; se prostitue pour/ sponsorise ; fournit une place de trottoir ; surveille ; fournit un logement, etc.). C'est à partir de ces actions que nous avons défini dix-neuf rôles (prostituée ; tutrice ; sponsor ; recruteur ; etc.). Nous dissocions les rôles strictement liés à l'activité criminelle d'exploitation des rôles liés à la migration (trolley/passeur ; coordonnateur du voyage) et au soutien associatif.

Les hypothèses associées et à vérifier sont les suivantes :

21. Pour rappel, on entend par "acteur", toute personne qui commet une action contribuant directement ou indirectement à l'exploitation.

— Les rôles sont cumulables chronologiquement et synchroniquement.

Nous avons à ce jour identifié le rôle de 201 individus. Le tableau plus bas liste le nombre d'invididus répertoriés selon le nombre de rôles endossés. Parmi les 129 individus endossant un rôle, on trouve 48 "prostituée" et 16 "sponsor".

Nombre de rôles endossés	1	2	3	4	5	6
Nombre d'invididus	129	37	16	13	4	2

— Il existe une différenciation sociale de la répartition des rôles :
  — Les rôles dans le réseau sont genrés.
  — La langue et/ou le dialecte parlés, le contexte culturel et plus généralement l'origine géographique semblent déterminants dans le rôle exercé. Par exemple, dans le cadre des réseaux nigérians, les personnes qui se prostituent ou qui "sponsorisent" parlent majoritairement Edo.

— Tous les acteurs continuent de faire de petites tâches. Les rôles ne sont pas strictement hiérarchisés mais leur cumul peut néanmoins indiquer une progression dans la "carrière" criminelle (Becker, 2012). Cette carrière est d'ailleurs associée à l'âge et à la position dans le cycle de vie.

A ce stade, on identifie des rôles plus importants que d'autres mais il reste difficile de tous les classer hiérarchiquement. Néanmoins, leur cumul est un indicateur pertinent d'une position forte au sein du réseau, lié certes à l'ancienneté dans le réseau mais également à la nature des rôles endossés alors. On observe ainsi un système d'ascension sociale qui fonctionne : le rôle de tutrice représente, par exemple, une évolution dans le réseau et préfigure un futur rôle de sponsor. Le soutien associatif, les rôles liés à la migration et les mères qui encouragent la prostitution sont des rôles périphériques et souvent uniques, ce qui semble indiquer qu'ils ne participent pas directement à l'exploitation. Ce pré-résultat est conforté par l'analyse qualitative des écoutes téléphoniques.

Par ailleurs, il faut rester prudent quant à la portée des données ; les 16 acteurs désignés par le seul rôle de sponsor ne peuvent être qualifiés de périphériques selon la même signification. La focale de l'enquête judiciaire ne révèle que peu d'informations sur ces individus. Nous faisons ici l'hypothèse qu'ils endossent d'autres rôles dans d' "autres réseaux". Rappelons ici que nous ne maitrisons pas les frontières de ce réseau, nous disposons de celles imposées par l'enquête judiciaire. Seule l'analyse de plusieurs dossiers de la même origine et procédant au même type d'exploitation pourrait révéler l'existence d'acteurs communs et permettre de dévoiler une part plus importante du réseau criminel.

— Il faut une combinaison forte du rôle et de la position structurale pour que l'individu soit considéré comme central au sein de l'activité criminelle du terme. En effet, le statut social entendu comme la position de l'individu dans la stratification sociale du réseau d'exploitation est lié à la combinaison du rôle et de la position structurale (mesure des centralités) dans le réseau.
  — Les rôles et positions sont concurrentiels.
  — La position structurale ne dépend pas du rôle et inversement.On peut avoir un rôle social important et une position structurale faible.
  — Le statut social est aussi lié aux rôles et positions structurales des individus avec lesquels la personne est en lien direct.

— Les individus avec les rôles et positions les plus faibles sont en lien direct avec des individus aux rôles et positions les plus forts.
— Le statut social dépend également de la position dans le cycle de vie : on observe une forte soumission aux ascendants (les filles sont soumises à leurs mères quel que soit le rôle de chacune) et de façon générale aux personnes plus âgées.

Toutes ces hypothèses doivent être traitées en considérant l'hétérogénéité du niveau des informations détenues : le statut social peut ainsi être renforcé par le fait que nous n'avons pas le même niveau d'information sur tous les acteurs.

De prime abord, la notion d'action n'est pas associée à un libre choix d'agir mais à une fonction bilatérale, réciproque, interdépendante et constitutive du réseau. Néanmoins, on peut se poser la question tant des déterminants sociaux et culturels que des logiques et motivations individuelles qui sous-tendent l'action. De la même manière, on peut se demander si la configuration sociale des réseaux criminels offre une liberté aux individus, entrevue ici comme la capacité d'agir sur le réseau d'interdépendance dans lequel ils sont inscrits (Elias, 1997).

3.1.2 L'analyse dyadique : type et contenu de la relation

Le second groupe d'hypothèses correspond au niveau meso de l'analyse. Un bon nombre de mesures peuvent être réalisées sur ce niveau d'analyse : le contenu, la nature, la fréquence, la force et l'orientation de la relation sont considérés comme des éléments explicatifs des liens et *a fortiori* de l'activité criminelle. La réciprocité et l'homophilie – être en lien avec des individus qui nous ressemblent socialement – peuvent également être interrogés.

Enfin, la modélisation par un modèle QAP – Quadratic assignment problem - permettra de dégager les effets d'attributs non structuraux sur l'existence d'un lien entre les acteurs [22].

Nous observons dans ce dossier 416 liens de réseaux, 255 liens financiers et 131 liens de connaissance. La complexité du travail entrepris tient pour partie au caractère multiplexe des liens identifiés. C'est ainsi que les liens identifiés précédemment peuvent se doubler de liens sexuels (41 identifiés) ou de liens de sang (107 identifiés).

3.1.3 L'analyse du réseau et du fonctionnement du système social

Enfin, le troisième groupe d'hypothèses, correspondant au niveau macro de l'analyse, permet d'appréhender le fonctionnement du groupe criminel ; il est entrevu et analysé ici selon les concepts de configuration [23], de réseau et de système social [24].

22. Ce modèle proposé par Krackhardt (1987) permet d'évacuer les caractéristiques structurales propres au réseau observé sans pour autant nier l'existence de corrélation entre les observations.

23. Elias considère la société comme un réseau d'interdépendances, un équilibre plus ou moins fluctuant de tensions. Cet équilibre est désigné par le concept de configuration qui renvoie à la forme que prend la structure à un moment donné. Reposant sur l'exemple du jeu où s'articulent concurrence et interdépendance, la configuration est donc en permanence reconstruite par les interactions des joueurs. Le concept oblige donc à penser la dynamique du réseau criminel et à adopter une démarche diachronique. Il permet également d'entrevoir une certaine représentation et concrétisation de la réalité et les règles qui sont en jeu (Elias, 1993).

24. Le système social, considéré ici selon Parsons est désigné par "une pluralité d'acteurs individuels inclus dans un processus d'interaction qui se déroule dans une situation affectée de propriétés physiques. Ces acteurs sont motivés selon une tendance à rechercher un "optimum de satisfaction", et leur situation est définie et médiatisée par un système de symboles, organisés par la culture à laquelle ils participent" (Parsons, 1955). La culture est définie ici comme un ensemble de valeurs et de symboles communs aux acteurs.

— Ce système repose sur une interdépendance fonctionnelle très forte de ses membres. Nous considérons ici que la dépendance réciproque des individus est constitutive du réseau criminel : les actions individuelles dépendent les unes des autres (Elias, 1997). Les relations entre les individus reposent principalement sur l'action réciproque et fonctionnelle sous-jacente.

— Il y a donc une forte division du travail (Durkheim, 2007) qui place chaque individu dans un rapport et des interactions fonctionnelles. La relation dyadique n'est donc pas tant liée à des rapports interpersonnels qu'à l'action réciproque au sein de l'activité criminelle.

— Cette interdépendance génère des contraintes sur l'individu. Le groupe criminel repose non pas sur une organisation pyramidale mais sur un équilibre des forces permettant ainsi de décrire la complexité et la dynamique inscrites dans ce réseau.

— Néanmoins, si le réseau est complexe – il contient des acteurs centraux, des rôles constitutifs et repose sur des actions spécialisées, tout acteur est en mesure de déstabiliser cet équilibre. Son mode de fonctionnement, dynamique, leur offre de solides ressorts. En effet, si on l'ampute de l'un de ses membres, il a la capacité de se régénérer et de faire évoluer les acteurs et rôles. Nous faisons ainsi l'hypothèse que certains agents qui le composent sont interchangeables : seule l'action, constitutive de l'activité criminelle, est stable. Autrement dit, le rôle prévaut sur l'individu.

— Le réseau est considéré par la sociologie des réseaux comme une structure sociale émergeant de l'ensemble des relations connectées entre elles (Degenne et Forsé, 2004). Il est de cette façon possible d'observer les régularités de forme de cette structure et d'en dégager une typologie (Bidart *et al.*, 2011). Ainsi, nous considérons que la forme du réseau, la forme que prennent les relations connectées, renvoie à un certain mode opératoire, un certain type de fonctionnement de réseau criminel.
  — Au sein d'un même type d'exploitation, les réseaux criminels n'ont pas le même mode opératoire selon l'origine géographique de ses acteurs. L'idée est donc de pouvoir comparer différents réseaux (bulgares, chinois, brésiliens, nigérians, etc.) et donc différentes formes de fonctionnement. La comparaison structurale des différentes formes de réseaux, permettra de vérifier cette hypothèse.
  — Le réseau criminel nigérian repose principalement sur des liens familiaux. On observe 107 liens de sang et 41 liens de couple au sein des 309 acteurs du réseau criminel.

— De la même façon, nous faisons l'hypothèse que les formes de réseaux et modes d'organisation associés dépendent du type d'exploitation et de la taille du réseau.

— Il existe une répartition géographique des réseaux de prostitution sur le territoire national en fonction des pays d'origine des prostituées. Le réseau nigérian est présent dans plusieurs lieux sur le territoire français sous la forme de sous-réseaux structuralement équivalents.

— Enfin, on considère que le réseau criminel est une société close. Un sous-groupe d'hypothèses y est associé :
  — Les individus ont peu de relations, voire de contacts, avec des individus hors du système d'exploitation.

— L'interconnaissance y est très forte et le contrôle social très important.
— Les relations sont multiplexes.
— Les relations familiales sont prégnantes.
— L'origine géographique des individus est homogène, y compris à l'échelle du "village" d'origine.

L'analyse structurale analysera notamment la taille du réseau – nombre de noeuds, nombre de liens entre les noeuds – et la connexité – le degré d'accessibilité, de dépendance, etc. La densité ne peut être mesurée car les frontières du réseau et l'existence des liens reposent sur la focale judiciaire du dossier.

Néanmoins, la structuration du réseau pourra être interrogée au regard de l'analyse des triades, clique et sous-groupes. Une modélisation à partir du modèle ERGM – Exponential Random Graph Models – permettra l'étude de la structure globale par une analyse à l'échelle du voisinage relationnel. Les caractéristiques structurales du réseau déterminent la probabilité d'existence d'un lien entre deux acteurs. Cette procédure met en lumière les interdépendances fonctionnelles du réseau lui-même, c'est à dire ses effets endogènes [25].

A terme, la structure des différents réseaux étudiés pourra être analysée conjointement à partir des analyses de l'équivalence structurale [26] des acteurs qui le compose, notamment par la procédure de *blockmodeling*.

3.2 Cartographie du réseau

Nous avons tout naturellement construit des cartes du réseau, à la fois à des fins de validation, et pour évoquer sa structure globale (tout en étant prudent sur les biais inévitables des représentations graphiques des réseaux). Les figures contenues dans cette section ont été obtenues à l'aide du logiciel de manipulation interactive de graphes Tulip [27] (Auber *et al.*, 2014, 2016).

Le réseau en Figure 3 présente l'ensemble des liens entre personnes sans égard au type des relations qui les lient ou à la date à laquelle une interaction est observée. Cette structure donne une idée d'ensemble des liens entre acteurs. Les sommets (acteurs du réseau) du graphe sont coloriés selon un gradient de bleu (du plus pâle au plus foncé) pour refléter le nombre de liens incidents à un acteur. La taille des sommets est calculée à partir de leur centralité d'intermédiarité. On observe que certains acteurs de fort degré n'ont toutefois qu'une faible centralité d'intermédiarité (ces deux paramètres de structure sont corrélés avec coefficient de Pearson de 0.88).

Il est plus intéressant, et c'est dans cette direction que nos travaux nous emmènent, de comparer les liens entre chaque "couche" du réseau (Burt et Scott, 1985; Battiston *et al.*, 2014; Renoust *et al.*, 2014). L'apport de la visualisation est de faciliter l'exploration des données en autorisant la formulation de requêtes dynamiques. Dans un contexte d'exploration interactive il devient possible, par exemple, de rechercher dans la couche constituée des liens financiers la position

25. "La distinction entre explication endogène et exogène de la présence liens sociaux est importante. Il est nécessaire de prendre en compte des tendances purement structurales de la formation de liens dans le but de faire les bonnes inférences concernant les effets des attributs des acteurs." (Lusher *et al.*, 2012, p. 27)

26. Des nœuds ont la même position dans un réseau, ils ont les mêmes liens et les mêmes non liens, des rôles sociaux sont ainsi identifiés et regroupés en blocs (ou positions) (Lorrain et White, 1971).

27. Le logiciel Tulip permet d'agencer simultanément plusieurs vues d'un même graphe, de sous-graphes, ou encore d'histogrammes et autres dispositifs graphiques dérivés de mesures ou d'attributs sur les graphes. Voir le site tulip.labri.fr

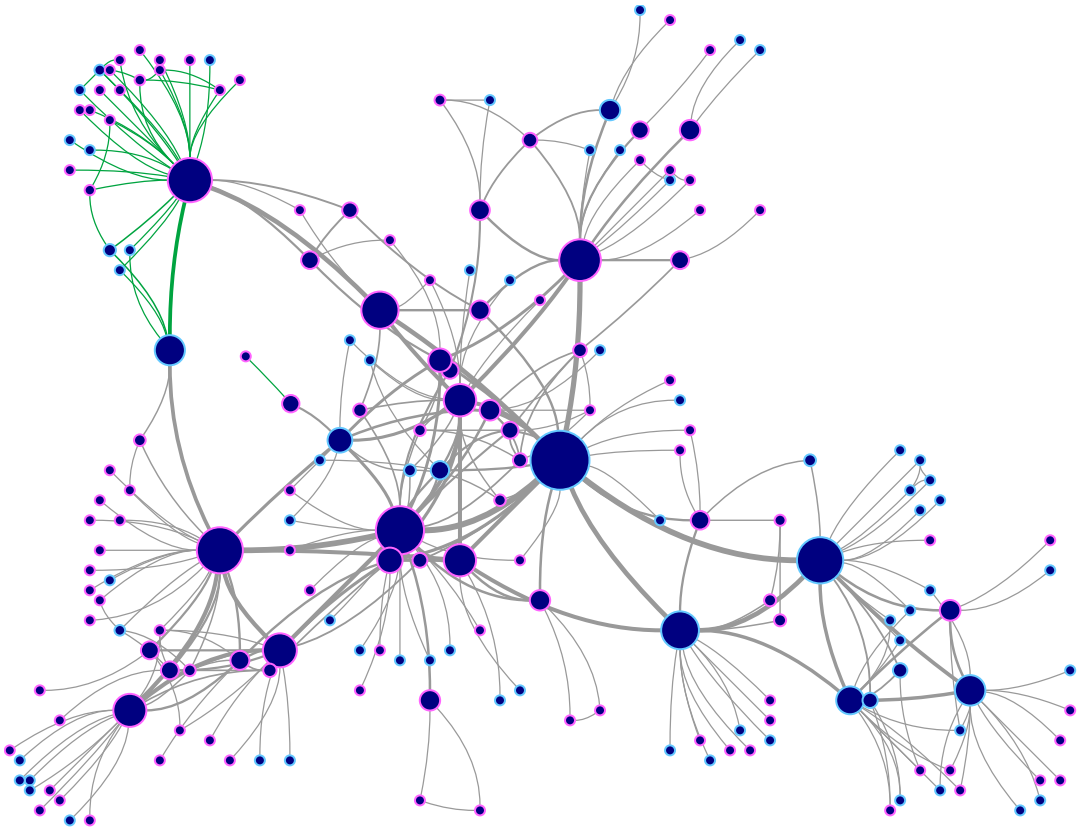


FIGURE 3 – Représentation schématique du réseau, tous types de liens confondus. Certains liens sont sélectionnés (vert) – voir les images suivantes.

d'acteurs d'un même cercle familial, ou encore de chercher à voir si les liens de connaissances sont orthogonaux aux tractations financières ou aux liens de réseau.

La Figure 4 montre la couche de "liens de reseau" extraite du graphe en Figure 3. La position des sommets dans chacune des couches reprend celle de la Figure 3, suivant un mécanisme d'héritages des propriétés calculées sur le graphe d'ensemble. La Figure 5 montre deux autres couches : "liens financiers" (haut) et "liens de sang" (bas). C'est à partir de ces vues synchronisées que l'analyste peut typiquement explorer le réseau de liens et formuler des requêtes sur les données.

Dans la vue des liens financiers (haut), une quinzaine de sommets et les liens entre ceux-ci ont été sélectionnés (la couleur verte marque la sélection dans la région supérieure gauche). Cette sélection opérée dans la couche "lien financiers" provoque la sélection des sommets correspondants dans les autres couches (s'ils y sont présents) ; sont ensuite sélectionnés les liens entre ces sommets dans chacune des couches. On constate naturellement une activité entre les acteurs sélectionnés sur la couche "liens de réseau". La sélection induite sur la couche "lien de sang" peut indiquer si les personnes impliquées ont par ailleurs des liens de sang.

Il ets possible de calculer un dessin particulier pour chacune de ces couches. La Figure 6 montre une portion du réseau de liens financiers tenant compte de l'orientation des flux (induit des échanges d'argent) allant du haut vers le bas. Ce type de dessin peut-être vu comme une requête dynamique effectuée sur le graphe sous-jacent : ce dessin hiérarchique donne une lecture des échanges permettant d'identifier les acteurs formant des points de "concentration" des flux, ou à l'inverse des points de "distribution" vers d'autres acteurs.

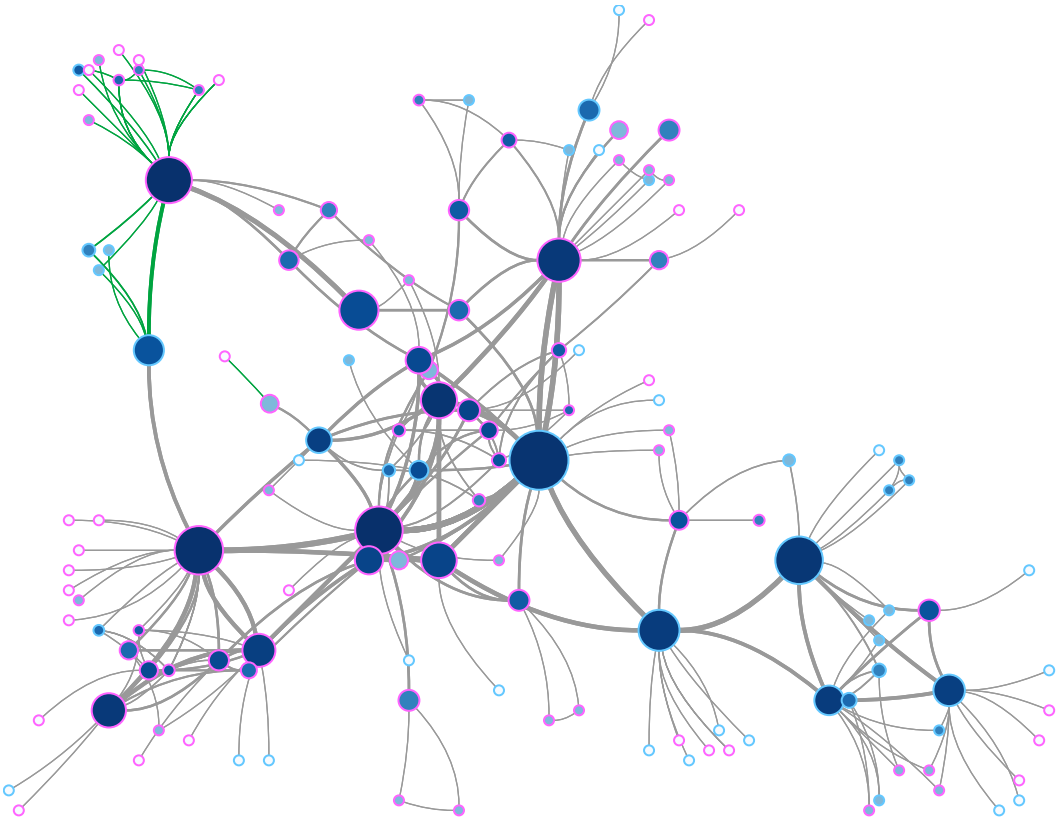


FIGURE 4 – Représentation des “liens de réseau” dont certains acteurs et liens ont été sélectionnés (partie supérieure gauche, liens marqués d’une couleur verte.

Des variables visuelles supplémentaires peuvent apporter des informations supplémentaires intégrées à même la représentation graphique de cette portion du réseau. Les bordures des sommets sont coloriés selon le sexe des acteurs : les bordures des sommets représentant des hommes y apparaissent en bleu clair alors que ceux représentant des femmes sont roses. De même, on pourrait utililser la forme des sommets pour distinguer les acteurs qui se prostituent, par exemple.

Ces figures illustrent parfaitement un scénario de fouille typique s’appuyant sur une cartographie interactive du réseau : combinaison des couches, dessin rendant compte d’une dynamique des liens, variables visuelles apportant une information sur les acteurs.

IV DISCUSSION ET TRAVAUX FUTURS

Cette section revient sur la méthodologie développée, sur ses limitations et sur la portée des analyses effectuées jusqu’à ce jour. Nous dessinons aussi certaines pistes à suivre pour consolider le volet informatique de nos travaux.

4.1 Discussion

L’utilisation du travail des enquêteurs pose donc un défi pour reconstruire un réseau dont on n’a qu’une image partielle de son activité. Nous comptons aussi faire appel à des approches probabilistes pour venir en aide à l’analyste et souligner l’absence de liens potentiels (Guimerà et Sales-Pardo, 2009). L’utilisation de modèles de génération aléatoire multi-couches existant (Méndez-Bermúdez *et al.*, 2017) est difficilement exploitable sans avoir auparavant caractériser les réseaux criminels que nous étudions.

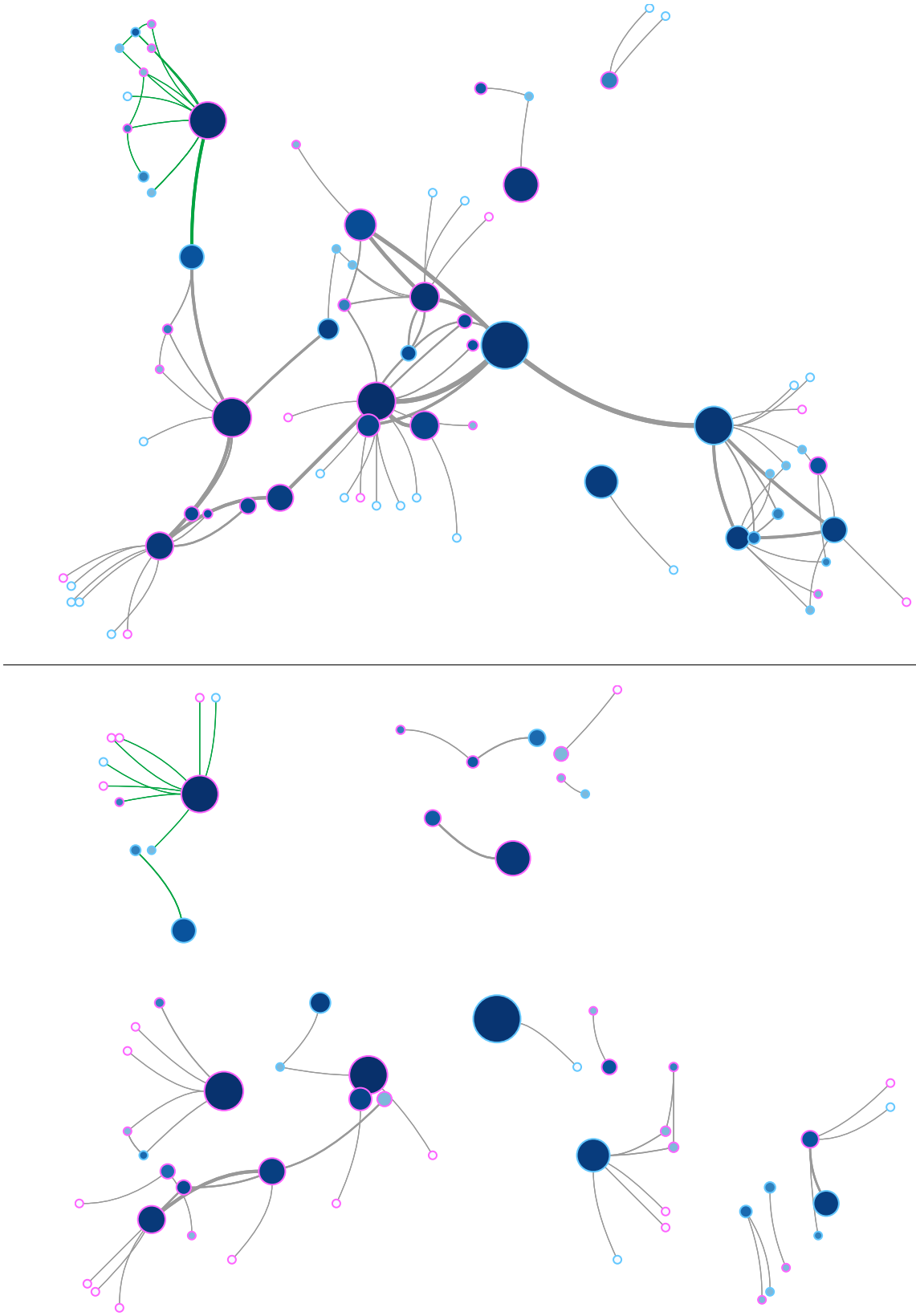


FIGURE 5 – Représentation simultanée de trois autres couches du réseau : “liens financiers” (haut) et “liens de sang” (bas). Les entités sélectionnées en figure 4 sont rapportées sur chacune des images (en vert aussi).

Volume : **2** - Année : **2017**, *DOI : 10.18713/JIMIS-300617-2-5*

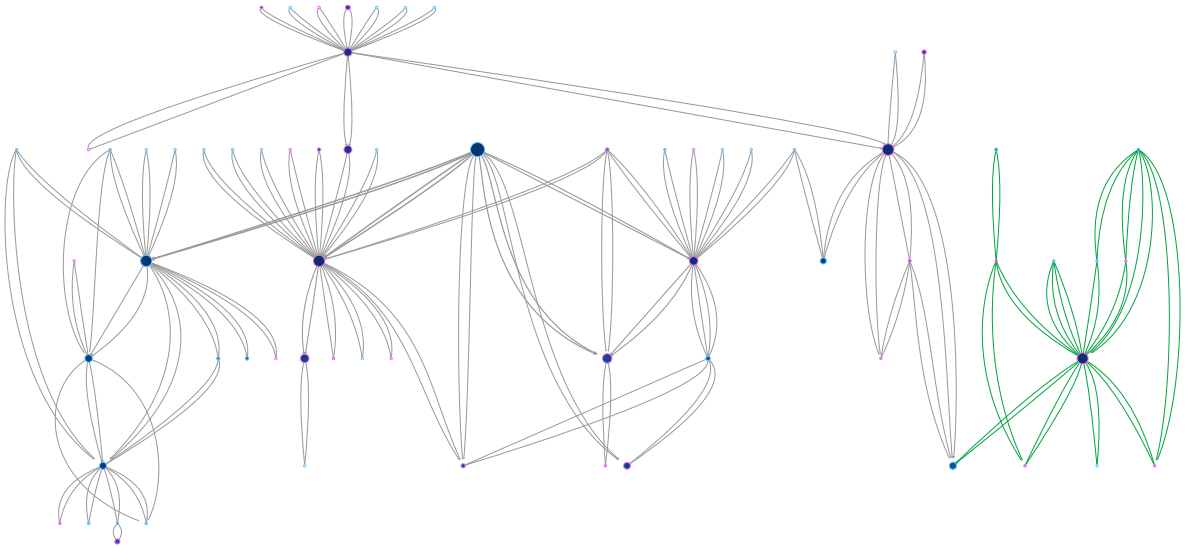


FIGURE 6 – Représentation hiérarchique (haut/bas) des liens financiers rendant compte de l'orientation des flux d'argent entre personnes – orienté vers le bas.

4.2 Contraintes de l'outillage informatique et pistes d'améliorations

Nous avons évoqué les limites du modèle de données relationnel utilisé pour construire une base de données stockant les informations extraites des dossiers judiciaires.

L'interface de saisie s'est vite révélée lourde d'utilisation, d'une part parce qu'elle était portée par une vision tabulaire, mais aussi parce qu'elle s'est construite à partir de composants existants qui se sont avérés plus contraignants que prévu. Or, le rôle de l'interface de saisie est un élément crucial de la méthodologie.

Bien que les premières analyses exploratoires et visuelles aient confirmé la pertinence du modèle de réseau multi-couches et qu'il ait été possible de stocker les attributs caractérisant chacune des couches, le modèle relationnel de données s'est encore une fois montré trop rigide. Cette rigidité se fait sentir en particulier lorsqu'il s'agit de formuler des requêtes mettant en jeu la structure de réseau ; le modèle relationnel se prête peu au calcul de voisinages dans les réseaux, par exemple.

Ces constats nous ont amené à envisager une refonte du modèle de données s'appuyant sur des moteurs de bases de données orientées "graphes". Nos premières expérimentations avec la technologie Neo4j (Webber, 2012; Robinson *et al.*, 2015) sont prometteuses. On peut voir une telle base de données comme un conteneur rassemblant des sommets et des liens entre ceux-ci. A chaque sommet et lien sont associés des *labels* qui en précisent le type, et auxquels s'ajoutent des propriétés (attributs). Ce modèle offre toute la flexibilité qui nous manque.

Cependant, les bases de données graphes nous privent de certains mécanismes propres aux bases de données relationnelles contrôlant les types de données des attributs. On y perd aussi la possibilité d'imposer des contraintes d'intégrité.

A l'évidence, on peut aussi envisager de faire reposer la saisie non pas sur des tableurs présentant de trop longues listes de personnes ou de liens, mais sur une représentation graphique des réseaux. Cela permettra de situer une entité – celle à laquelle on ajoute certaines informations – dans son contexte, ou son "voisinage réseau". De plus, il y aura ainsi continuité entre les représentations utilisées depuis la saisie jusqu'à l'analyse exploratoire et visuelle des données.

Les éléments qui précèdent mettent en évidence le potentiel de l'outil créé en termes de compréhension des réseaux criminels se livrant à des faits de traite des êtres humains. Si un travail important reste à accomplir pour pleinement exploiter notre méthodologie, nous sommes désormais en mesure de confirmer la faisabilité du projet initial, la richesse et le potentiel de la mise en oeuvre d'une démarche pluridisciplinaire.

Contributions des auteurs

B. Lavaud-Legendre, C. Plessard et G. Melançon ont assuré la rédaction de l'article alimenté par des discussions avec A. Laumond et B. Pinaud sur les aspects informatiques. B. Lavaud-Legendre a apporté le questionnement juridique et une expertise dans la lecture des dossiers. La conceptualisation et l'identification des variables suit une méthodologie apportée par C. Plessard. La collecte des données est attribuable à B. Lavaud-Legendre et H. Pohu. L'analyse de ce premier dossier est issue du regard croisé de B. Lavaud-Legendre et C. Plessard. Les réalisations logicielles sont le fruit du travail de A. Laumond, B. Pinaud et G. Melançon.

Références

Aghatise E. (2005). Réalités et cadre légal de la traite de nigérianes et d'européennes de l'est en italie. *Alternatives sud XII*(3), 135–164.

Aronowitz A. (2001). Smuggling and trafficking in human beings : The phenomenon, the markets that drive it and the organisations that promote it. *European Journal on Criminal Policy and Research 9*(2), 163–195. doi:10.1023/A:1011253129328.

Auber D., Archambault D., Bourqui R., Delest M., Dubois J., Pinaud B., Lambert A., Mary P., Mathiaut M., Melançon G. (2014). Tulip III. In R. Alhajj et J. Rokne (Eds.), *Encyclopedia of Social Network Analysis and Mining*, pp. 2216–2240. New York : Springer. doi:10.1007/978-1-4614-6170-8_315.

Auber D., Bourqui R., Delest M., Lambert A., Mary P., Melançon G., Pinaud B., Renoust B., Vallet J. (2016). Tulip 4. Technical report, Université de Bordeaux, CNRS UMR 5800 LaBRI. URL : https://hal.archives-ouvertes.fr/hal-01359308.

Barabási A. L. (2011). The network takeover. *Nature Physics 8*(1), 14. doi:10.1038/nphys2188.

Battiston F., Nicosia V., Latora V. (2014). Structural measures for multiplex networks. *Physical Review E 89*(3), 032804. doi:10.1103/PhysRevE.89.032804.

Becker H. (2012). *Outsiders : Etudes de sociologie de la déviance*. Editions Métailié.

Besnard P., Boudon R., Cherkaoui M., Lécuyer B.-P. (1999). *Dictionnaire de sociologie*. Larousse.

Bidart C., Degenne A., Grossetti M. (2011). *La vie en réseau : dynamique des relations sociales*. Presses Universitaires de France.

Bourdieu P. (1986). L'illusion biographique. *Actes de la recherche en sciences sociales 62/63*, 69–72.

Bright D., Greenhill C., Reynolds M., Ritter A., Morselli C. (2014). The use of actor attributes and centrality measures to identify key actors : A case study of a drug trafficking network. *Journal of Contemporary Criminal Justice 31*(3), 262–278. doi:10.1177/1043986214553378.

Brown T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd Edition ed.). The Guilford Press.

Burt R., Scott T. (1985). Relation content in multiple networks. *Social Science Research 14*, 287–308.

Calderoni F. (2012). The structure of drug trafficking mafias : the 'Ndrangheta and cocaine. *Crime, Law and Social Change 58*(3), 321–349. doi:10.1007/s10611-012-9387-9.

Calderoni F. (2014). Social network analysis of organized criminal groups. In G. Bruinsma et D. Weisburd (Eds.), *Encyclopedia of Criminology and Criminal Justice*, pp. 4972–4981. New York : Springer. doi:10.1007/978-1-4614-5690-2_239.

Campana P. (2016). The structure of human trafficking : lifting the bonnet on a nigerian transnational network. *British Journal of Criminology 56*(1), 68–86. doi:10.1093/bjc/azv027.

Carrington P. J. (2014). Crime and social network analysis. In J. P. Scott et P. J. Carrington (Eds.), *The sage handbook of social network analysis*, pp. 236–255. SAGE Publication. doi:10.4135/9781446294413.n17.

Colombié T., Laman N., Schiray M. (2001). *Les acteurs du grand banditisme français au sein des économies souterraines liées au trafic de drogue*. Centre international de recherche sur l'environnement et le développement. URL : https://criminocorpus.org/fr/outils/bibliographie/consultation/ouvrages/131047/.

Crozier M., Friedberg E. (1992). *L'acteur et Le Système : Les Contraintes de L'action Collective*. Seuil.

Dale A. (1993). Le rôle de l'analyse secondaire dans la recherche en sciences sociales. *Sociétés contemporaines 14*(1), 7–21.

de Domenico M., Porter M. A., Arenas A. (2015). MuxViz : a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks 3*(2), 159–176. doi:10.1093/comnet/cnu038.

Degenne A., Forsé M. (2004). *Les réseaux sociaux*. Armand Colin.

Delley J.-D., Flueckiger A. (2005). La légistique : une élaboration méthodique de la législation. In R. Drago (Ed.), *Confection de la loi*, pp. 83–96. Paris : Presses Universitaires de France.

Durkheim E. (2007). *De la division du travail social*. Presses Universitaires de France.

Elias N. (1993). *Qu'est-ce que la sociologie ?* Editions de l'Aube.

Elias N. (1997). *La société des individus*. Fayard.

Fijnaut C., Bovenkerk F., Bruinsma G., van de Bunt H. (1998). *Organized Crime in the Netherlands*. Kluwer.

Granovetter M. (1985). Economic action and social structure : The problem of embeddedness. *American Journal of Sociology 91*(3), 481–510.

Grossetti M. (2009). Qu'est-ce qu'une relation sociale ? un ensemble de médiations dyadiques. *REDES - Revista Hispana Para El Análisis de Redes Sociales 16*(2), 44–62.

Guimerà R., Sales-Pardo M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences 106*(52), 22073–22078. doi:10.1073/pnas.0908366106.

Kivelä M., Arenas A., Barthelemy M., Gleeson J. P., Moreno Y., Porter M. A. (2014). Multilayer networks. *Journal of Complex Networks 2*(3), 203 – 271. doi:10.1093/comnet/cnu016.

Krackhardt D. (1987). Cognitive social structures. *Social Networks 9*(2), 109–134. doi:10.1016/0378-8733(87)90009-8.

Lavaud-Legendre B., Quattoni B. (2013). Désir migratoire, emprise et traite des êtres humains. In L.-L. B. (Ed.), *Prostitution nigériane*, pp. 61–92. Karthala.

Lazarsfeld P. F. (1965). Des concepts aux indices empiriques. In Mouton (Ed.), *Le vocabulaire des sciences sociales : concepts et indices*, Méthodes de la sociologie, Chapter 1, pp. 27–36. Mouton.

Lorrain F., White H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology 1*(1), 49–80. doi:10.1080/0022250X.1971.9989788.

Lusher D., Koskinen J., Robins G. (2012). *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. Cambridge University Press.

Malm A., Bichler G. (2011). Networks of collaborating criminals : Assessing the structural vulnerability of drug markets. *Journal of Research in Crime and Delinquency 48*(2), 271–297. doi:10.1177/0022427810391535.

Mancuso M. (2014). Not all madams have a central role : analysis of a nigerian sex trafficking network. *Trends Organ Crim 17*(1-2), 66–88. doi:10.1007/s12117-013-9199-z.

March J.-G., Simon H.-A. (1999). *Les organisations : problèmes psychosociologiques*. Dunod.

Mercklé P. (2004). *Sociologie des réseaux sociaux*. La Découverte.

Meyer M., Sedlmair M., Munzner T. (2012). The four-level nested model revisited : Blocks and guidelines. In *Workshop on BEyond time and errors : novel evaLuation methods for Information Visualization (BELIV)*. doi:10.1145/2442576.2442587.

Mitchell J. C. (1969). *Social Networks in Urban Situations : Analyses of Personal Relationships in Central African Towns*. Manchester University Press.

Morselli C. (2010). Assessing vulnerable and strategic positions in a criminal network. *Journal of Contemporary Criminal Justice 26*(4), 382–392. doi:10.1177/1043986210377105.

Munzner T. (2009). A nested process model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics 15*, 921–928. doi:10.1109/TVCG.2009.111.

Méndez-Bermúdez J. A., Ferraz de Arruda G., Rodrigues F. A., Moreno Y. (2017). Scaling properties of multilayer random networks. *Physical Review E 96*(1), 012307. doi:10.1103/PhysRevE.96.012307.

Newman M., Barabási A.-L., Watts J. (2006). *The structure and dynamics of networks*. Princeton Studies in Complexity. Princeton University Press.

Osusky M. (2007). Utilization of multilayer network data of team for sociomapping analysis. In *XXVII Sunbelt Conference*. INSNA International Network for Social Network Analysis.

Parsons T. (1955). *Eléments pour une sociologie de l'action*. Plon.

Renoust B., Melançon G., Viaud M.-L. (2014). Entanglement in multiplex networks : Understanding group cohesion in homophily networks. In R. Missaoui et I. Sarr (Eds.), *Social Network Analysis - Community Detection and Evolution*, Lecture Notes in Social Networks, Chapter 5, pp. 89–117. Springer. doi:10.1007/978-3-319-12188-8_5.

Robinson I., Webber J., Elfrem E. (2015). *Graph Databases* (2nd ed.). O'Reilly.

Rossy Q. (2011). *Méthodes de visualisation en analyse criminelle : approche générale de conception des schémas relationnels et développement d'un catalogue de patterns*. Phd thesis, Université de Lausanne, Faculté de droit et des sciences criminelles. URL : https://www.unil.ch/esc/files/live/sites/esc/files/shared/The_se_Rossy.pdf.

Rossy Q. (2016). La visualisation relationnelle au service de l'enquête criminelle. In C. Morselli (Ed.), *Les réseaux criminels*. Presses de l'Université de Montréal.

Salt J. (2000). Trafficking and human smuggling : A european perspective. *International Migration 38*(3), 31–56. doi:10.1111/1468-2435.00114.

Salt J., Stein J. (1997). Migration as a business : The case of trafficking. *International Migration 35*(4), 467–94. doi:10.1111/1468-2435.00023.

Sardi M., Froidevaux D. (2003). Le monde de la nuit : milieu de la prostitution, affaires et 'crime organisé'. Technical Report 4040-054324, PNR40 "Violence et criminalité organisée", FNRS. doi:10.13140/RG.2.1.5068.8082.

Silverman D. (1973). *La théorie des organisations*. Dunod.

Strauss A. (1992). *Miroirs et masques. Une introduction à l'interactionnisme*. Métailié.

Thomas J. J., Cook K. A. (2006). *Illuminating the Path : The Research and Development Agenda for Visual Analytics*. IEEE Computer Society.

Tukey J. W. (1977). *Exploratory Data Analysis*. Reading, MA : Addison-Wesley.

van Dijk J. J. M., van der Knaap M., Aebi M. F., Campistol C. (2014). Counting what counts ; tools for the validation and utilization of eu statistics on human trafficking. Report, INTERVICT/T/Universitat Autònoma de Barcelona.

Webber J. (2012). A programmatic introduction to Neo4j. In *3rd annual conference on Systems, programming, and applications : software for humanity*, pp. 217–218. ACM. doi:10.1145/2384716.2384777.

Weitzer R. (2014). New directions in research on human trafficking. *The annals of the american academy of political and social science 653*(1), 6–24. doi:10.1177/0002716214521562.

Xu J., Chen H. (2005). Criminal network analysis and visualization. *Communications of the ACM 48*(6), 100–107. doi:10.1145/1064830.1064834.