Mind Soc (2016) 15:155–167
DOI 10.1007/s11299-015-0172-1

Information-driven network analysis: evolving the "complex networks" paradigm

Remo Pareschi[1] · **Francesca Arcelli Fontana**[2]

Received: 3 February 2015 / Accepted: 30 April 2015 / Published online: 11 July 2015

Abstract Network analysis views complex systems as networks with well-defined structural properties that account for their complexity. These characteristics, which include scale-free behavior, small worlds and communities, are not to be found in networks such as random graphs and lattices that do not correspond to complex systems. They provide therefore a robust ground for claiming the existence of "complex networks" as a non-trivial subset of networks. The theory of complex networks has thus been successful in making systematically explicit relevant marks of complexity in the form of structural properties, and this success is at the root of its current popularity. Much less systematic has been, on the other hand, the definition of the properties of the building components of complex networks. The obvious assumption is that these components must be nodes and links. Generally, however, the internal structure of nodes is not taken into account, and links are serendipitously identified by the perspective with which one looks at the network to be analyzed. For instance, if the nodes are Web pages that contain information about scientific papers, one point of view will match the relevant links with hyperlinks to similar Web pages, and another with citations of other articles. We intend to contribute here a systematic approach to the identification of the components of a complex network that is based on information theory. The approach hinges on some recent results arising from the convergence between the theory of complex networks and probabilistic techniques for content mining. At its core there is the idea that nodes in a complex network correspond to basic information units from which links

✉ Remo Pareschi
remo.pareschi@unimol.it

Francesca Arcelli Fontana
arcelli@disco.unimib.it

[1] Department of Bioscience and Territory, University of Molise, Pesche, IS, Italy

[2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

are extracted via methods of machine learning. Hence the links themselves are viewed as emergent properties, similarly to the broader structural properties mentioned above. Indeed, beside rounding up the theory, this approach based on learning has clear practical benefits, in that it makes networks emerge from arbitrary information domains. We provide examples and applications in a variety of contexts, starting from an information-theoretic reconstruction of the well-known distinction between "strong links" and "weak links" and then delving into specific applications such as business process management and analysis of policy making.

Keywords Complex systems · Complex networks · Information theory · Probabilistic topic models · Business process management · Policy analysis

1 Introduction

The study of complex systems has recently received a vigorous surge from the evidence that, in addition to those among such systems that have evolved in nature as well as in society during a more or less remote past, many of the most relevant developments made possible by the coming of age of a digital society bear the marks of complexity. These developments include lower-level infrastructures for the transfer of digital information like the Internet as well as strata located at higher levels, like the World-wide Web, social networks and the blogosphere, where human users play directly a key role in the creation and exchange of information. This aspect throws a new light on the issue of the relationship between information and complexity, namely whether this relationship, that plays such a crucial part in these recent systems because of their inherently information-driven nature, should not be radically revised for the purposes of the study of complexity in general.

Of course the relationship between complexity and information builds itself on a long and important tradition, rich in substantial results which have had both theoretical and practical implications. Specifically, such contributions can be found in Shannon entropy and Kolmogorov complexity where the complexity of an information unit is reconstructed in terms of the complexity of the minimal algorithm that can reproduce it without loss of information (Li and Vitányi 2008). Shannon entropy and Kolmogorov complexity have however very little to say about situations where the system to be analyzed is composed by several connected units. By contrast, it is here that the application of techniques and methodologies of network analysis to the field of complex systems has given some important contributions (Newman et al. 2006), and has done so by focusing on topology rather than on information. Therefore, so as to account for complexity, a number of well-defined structural properties of complex networks has been identified, such as scale-free behavior, small worlds and communities. Much less systematic has been, on the other hand, the definition of the properties of the building components of complex networks. The obvious assumption is that these components must be nodes and links. Generally, however, the internal structure of nodes is not taken into account, and links are serendipitously identified by the perspective with which one looks at the network to be analyzed. For instance, if the nodes are Web pages that contain information about

scientific articles, one point of view will match the relevant links with hyperlinks to similar Web pages, and another with citations of other articles. Theoretical incompleteness aside, this state of affairs severely limits the practical applicability of network analysis to real cases of complexity, because of the arbitrariness in setting out the initial premises that it entails, and the reservations that inevitably ensue on the actual empirical validity and effective significance of the results.

We show here how a systematic approach to the identification of the components of a complex network can be provided precisely by bringing information back into the picture, thus paving the road to bringing together two theoretical views on complexity as well as fulfilling a practical need. The approach hinges on the exploitation of probabilistic techniques for content mining and information-retrieval. At its core there is the idea that nodes in a complex network correspond to basic information units from which links are extracted via methods of machine learning. Hence the links themselves are viewed as emergent properties, similarly to the broader structural properties mentioned above.

The intent of the paper is conceptual and methodological, and relies on the technical background and the formal results of Rossetti et al. (2014). It is structured for its remaining parts as follows: in Sect. 2 we provide an information-theoretic framework for complex networks; in Sect. 3 we demonstrate how such framework successfully fulfills the objective of making complexity phenomena systematically emerge from an information substratum in the form of the well-known structural properties studied in network theory, and we especially focus on properties of special interest for social network analysis such as communities and the distinction between weak links and strong links; in Sect. 4 we go into specific case studies and applications in such domains as business process management and policy analysis. Section 5 concludes the paper.

2 Radical textualism (and empiricism)

Our primary conceptual and methodological assumption is that nodes of a network can be coded as texts. If these nodes are meant to be textual objects or documents of some sort, such as Web pages, articles from a scientific corpus, elements of a code of law, components of a software library etc., then this aspect follows obviously and immediately. But a difference and indeed an inconsistency, or, on the contrary, an excess of reductionism, are defects that apparently can be attributed to this approach whenever the network scopes over nodes that are not texts—for instance, if the nodes are people, animal species, or logistic ports, just to mention a few cases that have been the focus of recent investigations in the application of network theory. Nonetheless, when confronted with these arguments, we stick to the claim that it is totally acceptable to view even nodes of this kind as texts. Our reasoning is based on the reconstruction of the initial step of the process of network analyst, namely setting up the network itself. This step must be carried out through the identification of the links that build up the network, and to this aim use must be made of implicit or explicit textual descriptions of the nodes. For instance, in setting up a network corresponding to a “Jungle Book” ecosystem, texts will be used which state that

tigers are apex predators, cobras are generally apex predators even if they are sometimes victims of attacks from mongooses, wild dogs are social predators, blackbucks are a favored food of tigers etc., thus accessing the necessary information for establishing predator–prey links. Hence, we can generalize this point by saying that all the information about nodes of a network is or can be documented in texts, and that for each node there is a text that contains information from which links can be identified and extracted.

As simple and straightforward as it is, this premise is the keystone of all, since it provides a stringent motivation for an information-driven overhaul of the state of the art of complex networks as well as points the way for its satisfactory implementation. The motivation stems both from the opportunity of automating and from the necessity of making less arbitrary the identification of the links that boots up network analysis. The road to a satisfactory implementation pursues the possibility to apply techniques of text analysis and machine learning to the nodes of the network, that are now viewed as texts, so as to automatically generate and extract links through an empirically grounded procedure.

2.1 Generative network modeling

Thus, we are working under a strongly generative perspective. The effect of this is that not only links are systematically generated, but that structural properties corresponding to diverse aspects of complexity emerge directly during the process of network generation, as opposed to being detected afterwards as traditionally takes place. To see how all this happens we introduce now some fundamental concepts.

Since nodes are texts, one immediate way of generating networks on the basis of the information that they contain (and hence by strictly empirical criteria) is by grouping them according to the topics that they share; in other words, nodes that talk about the same things are linked together and therefore generate a network. Of course, given that a node might correspond to a text that talks about multiple topics, there might be multiple networks it belongs to, just as someone who is a lawyer and is also a keen golfer fits naturally within two human networks, the one of lawyers and the one of golfers.

Fortunately, this clear and intuitive conceptualization of the relationship between nodes as units of textual information, topics and networks, receives direct support by the proven formal and computational background given by Probabilistic Topic Modeling (PTM) (Blei 2012). PTM was developed in computer science in the last decade with the purpose of making feasible the analysis on probabilistic bases of large volumes of textual data and, by virtue of our simple move, can be transferred tout court to the study of complex networks. The main purpose of PTM algorithms is the analysis of words in natural language texts in order to discover topics represented by sorted lists of words. For instance, Fig. 1 shows 4 out of 300 topics extracted from the TASA corpus (Steyvers and Griffiths 2007). It is easy to see that words in the four topics are related to each other and can be considered as consistent themes. Furthermore, PTM is also able to provide topic proportions for each document, which is very useful to understand which topics a document is about. Seen in the context of machine learning and knowledge discovery, PTM-based text

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Fig. 1 Topics from the TASA corpus

mining approaches aim to get the best of both worlds, by providing richly structured representations of the knowledge derived from the empirical validation of "Big Data" processing. Hence they improve both on traditional symbolic approaches, that lack data validation, and on connectionist approaches, that lack capability to represent knowledge (Tenenbaum et al. 2011).

The gist is that we use PTM to group nodes into networks that correspond to topics, and consequently we generate the links connecting the nodes. Clearly all nodes within such networks will be connected to each other (in graph-theoretic terms they will be cliques), and thus will be by definition, according to a popular terminology, small worlds and, within the contexts of larger networks, communities, namely denser regions of the networks. We have in this way addressed one half of the question, the other one being related to the generation of full complex networks, where communities and small worlds are only a part of the picture. In other words, can we generate, using our information-theoretic approach, full-fledged complex networks, where there are also links that bridge "distant" regions of the networks? Obviously this is a result that is needed and expected since, starting from typical populations of individual nodes that characterize complex phenomena, we want precisely to generate complex networks in their full generality.

As was previously outlined (Rossetti et al. 2014), it turns out that this is feasible. We can indeed introduce a notion of probabilistic transition from one information unit to the other, derived from the probabilistic topic distribution associated with the units themselves. This kind of one-step probabilistic transitions is formally rooted into the well-known framework of Markov chains and lets us spread the network, so to say, so as to catch within it all the communities associated with the diverse topics by establishing links that connect members belonging to different communities and thus cut across community boundaries. The same technique can be applied at the

higher-level to the topics themselves. In this way we can leverage empirical corroboration to define an ontological subsumption relation between topics: for example, a 0.60 probability of transitioning from a topic that groups Web pages on some recent terrorist attacks to one that groups Web pages related to railway stations and coach stations boils down to a relevant probability that such attacks may occur in a logistic context of this type. Notice also how this correlation indeed extracts new information and knowledge which could hardly be obtained through the traditional lexicographic approach of ontology compilation. In fact, in terms of settled linguistic knowledge, terrorist activity and transport logistics are very distant concepts, that are brought together by real-world dynamics happening so fast that are very difficult to predict under the lens of lexicographic evolution.

Another point of view on the different types of links that are brought together, respectively, via topic clustering and via one-step probabilistic transitions is that they define a generative framework for the well-known distinction grounded on social behavior between strong links and weak links (Granovetter 1983). Citing *verbatim* from the Wikipedia entry on interpersonal ties http://en.wikipedia.org/wiki/Interpersonal_ties "In mathematical sociology, *interpersonal ties* are defined as information-carrying connections between people. Interpersonal ties, generally, come in three varieties: *strong*, *weak*, or *absent*. Weak social ties, it is argued, are responsible for the majority of the embeddedness and structure of social networks in society as well as the transmission of information through these networks. Specifically, more novel information flows to individuals through weak rather than strong ties. Because our close friends tend to move in the same circles that we do, the information they receive overlaps considerably with what we already know. Acquaintances, by contrast, know people that we do not, and thus receive more novel information." This situation is clearly reconstructed with our approach in a formally precise information-theoretic fashion through the building up of strong links via topic clustering and then evolving the minimal social networks thus defined into full-fledged social networks via weak links generated as one-step probabilistic transitions. Even more, by applying probabilistic transitioning to topics, we can generate a conceptual counterpart, or superstructure, of the underlying social network.

3 A case study: generating a social network of terroristic activities

Terrorism is certainly not the most cheerful domain from which to extract social networks but is just as certainly one of the most relevant, given the recent insurgence of terrorist activities that derive from decentralized mechanisms of cooptation among groups and individuals. What we provide here as an example is a social network that maps out connections among all the relevant information units in past terrorist activities, including reports on events and places where they have happened.

Figure 2 shows the set of generated topics with their associated probabilistic transitions, while Fig. 3 provides a snapshot of the network with links that cut across topics and act as weak ties. Note that in Fig. 2 only the words of greater weight within the topics have been highlighted, and these same words are printed out with different sizes to reflect their different weights.

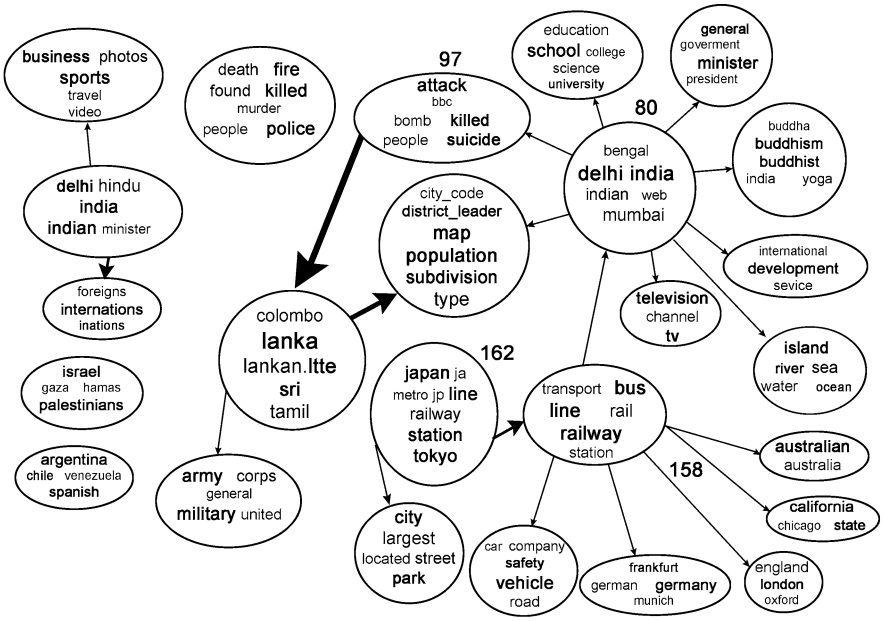

Fig. 2 Topic-topic network

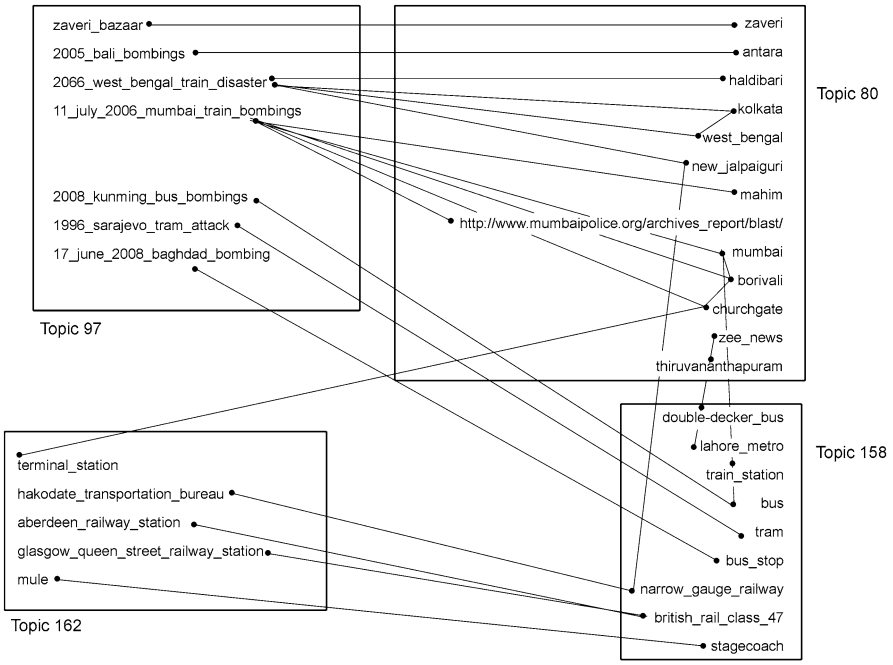

Fig. 3 Links crossing topic boundaries

Figure 2 identify conceptual relationships within the "information society" of terrorism, such as the fact that topic 162 about a terrorist attack on the Tokyo subway line is related to a more general topic, numbered 158, about urban public transportation. The object–object links of Fig. 3 explode these views into similar relationships, but at a higher resolution: for instance a Web page related to the train bombings in Mumbai in 2006 in topic 97 is linked to a variety of Web pages in topic 90 that are specific to the city of Mumbai, such as Web pages of its railway stations. The information society of terrorism is in this way constructed bottom-up and inside-out, by connecting through weak links topics with topics and objects from one topic to another.

A question that naturally arises is how this generative approach based on semantics and driven by information compares with the more traditional approaches to network analysis that, by assuming a basic structural configuration (in the form of links) of the network to be analyzed, proceed accordingly to isolate specific properties such as the presence of communities. Clearly, the generative and information-driven approach is more general, in that it can operate in the absence of explicit links and can generate links by leveraging the full semantics made available by the information contained within the nodes. However, what about when "natural" explicit links are clearly available, as is the case with the hyperlinks that connect together Web pages? In this case, it might seem reasonable to apply algorithms for the detection of structural properties that skip the level of semantic analysis altogether by targeting directly the graph structure, such as those proposed in Harel and Koren (2001), Newman and Girvan (2004) and Rosvall and Bergstrom (2008). However, results from experiments reported in, involving among others the data set corresponding to the information society of terrorism described in this section, show that even in this case our semantic approach fundamentally outperforms the graph-oriented algorithms. Aside from mere performance issues, where it fares well anyway, semantic generation can indeed do better at detecting relevant structural properties than graph analysis even when the syntax of content composition supports the definition of links pointing to other contents, as is the case with the World-wide Web and hypertexts in general. The reason is obvious: in the case of very large hypertexts like the WWW contents are generated through a variety of content providers who do not need to coordinate and may not be aware of each other, even if they produce related contents, and may anyway be unwilling to coordinate and cooperate. They may therefore produce contents that are related but are not linked. Our approach lets such "missing links" freely emerge, and can consequently identify more communities, small worlds and other structural properties with respect to graph-oriented approaches that are limited to the available explicit links.

4 Two application domains: business process management and policy analysis

Aside from the generalist context of public domain large content corpora, such as the WWW, that we have exemplified so far, it is interesting to explore the effectiveness of the semantic generative approach to network analysis in the context

of two domain-specific application domains: business process management and policy analysis.

4.1 Business process management

Business processes have been themselves seen in the past from a strictly structural point of view, as sequences of actions to be executed on the basis of rigid and pre-defined information flows, which bind together agents with the appropriate capabilities for the performance of different types of tasks and the tasks associated with execution of the given actions. However, more recently the stress has been on business processes capable of coping with the flexible organizations of the digital and post-industrial age. Hence there is need of making relationships between agents, tasks and actions emerge in a much more flexible way, by filtering and matching information profiles in a way that closely resembles the generative approach to complex networks described above.

4.1.1 Software development and software ecosystems

Probabilistic Topic Models have indeed been used for the purpose of modeling software systems as complex ecosystems and hence for studying the evolution of software, with the aim to design monitoring techniques suitable for the management of very large software development processes, lasting over extended timeframes and involving large teams of developers often distanced from each other both in time and space. This is a very mature area, which pre-dates the recent insights and developments that relate network topology with information retrieval, and nevertheless show very effectively the practical usefulness and explanatory power of information-driven techniques in the management of complex systems developed and maintained through a variety of business processes. We provide here a brief survey of the state of the art.

Given that we are talking about systems existing and evolving in a timeframe, time plays a relevant role and one important inspiration has been a more general work by Hall et al. (2008), where topic evolution in time is used to track disciplinary evolution of ideas. Linstead et al. (2008) have applied the Hall et al. approach to a system's source code release history to discover topics and their evolution within the source code. Thomas et al. (2014) extend their work by starting from the assumption that "topic evolution provides a unique opportunity for automatically monitoring a project's source code over time". If the monitoring activity identifies topics that are more and more scattered on the system, probably these can give hints about the parts of code that need to be better investigated and refactored. Through such monitoring activity it is also possible to address issues related to checking the need of including or removing some specific functionality in the system.

Rama et al. (2008) use topic models to automatically mine business topics from a snapshot of a software system and from its version history. Thomas et al. (2010) exploit and test these results with respect to the task of determining whether the discovered topic evolutions represent an accurate description of the actual change activity in the source code. They characterize and define topic evolution in this

contest and analyze the causes that make topics evolve for the purpose of studying the evolution of software.

Kuhn et al. (2007) introduced semantic clustering, based on Latent Semantic Indexing (LSI), an extensively studied algorithmic variant of Probabilistic Topic Models, to group source code documents. The documents are clustered on the basis of their similarity into semantic clusters, corresponding to the implementation of similar functionalities. Kawaguchi et al. (2006) introduced a tool named MUDABlue that also uses LSI to automatically classify software systems into categories based on the identifiers in the system. This is useful for browsing large, unlabeled collections of software. Tian et al. (2009) modified the MUDABlue approach to employ another algorithmic variant, Latent Dirichlet Allocation (LDA), instead of LSI, so as to consider comments as well as identifiers. The two approaches achieve comparable performance.

Bavota et al. (2014a) address software modularization problems. They partition underlying latent topics in classes and packages and use structural dependencies to recommend refactoring operations aiming at moving classes to more suitable packages. The topics are acquired via Relational Topic Models (RTM) (Chang and Blei 2009), a probabilistic topic modeling technique, which is elsewhere exploited to capture also coupling among classes in object-oriented software systems (Gethers and Poshyvanyk 2010). They developed a tool to support software re-modularization called R3 (Rational Refactoring via RTM) which was evaluated it in two empirical studies. An interesting feature of R3 is that it provides support to software developers in evaluating the goodness of suggested refactoring operations by generating explanations for the suggested operations by using topic analysis. In another work, Bavota et al. (2014b) describe yet another tool, called Methodbook, used as a recommendation system for method refactoring operations with the aim to improve the design quality of the systems as captured by quality metrics.

Panichella et al. (2013) propose LDA-GA, an approach based on genetic algorithms that determines the near-optimal configuration for LDA in the context of three important software engineering tasks, namely traceability link recovery, feature location, and software artifact labeling. The combination of two approaches, both based on machine learning, namely genetic algorithms and LDA/PTM, is yet another instance of the effectiveness and of the flexibility of the learning paradigm in coping with real-world problems in systems and processes that are inherently bound towards complexity.

4.1.2 Monitoring of open-world processes

A more recent application, which merges network generation as presented here with the seminal work of Hall et al. (2008) about evolution of topics over time, is the monitoring of open-world business processes, namely processes that involve multiple organizations and a large variety of stakeholders and are strongly influenced by external factors. Such processes exist in reality even if they most often lack a complete definition, precisely because they arise from the combination of sub-processes owned by single organizations. As matter of fact, because of their width and complexity, they are crucial processes for measuring the effectiveness of

large social and economic ecosystems in the fulfilment of their aims and purposes, such as the well-being of citizens, customer satisfaction, equilibrium among stakeholders etc.

In Pareschi et al. (2014) this is illustrated through a case study on the deployment of labor law, with processes that involve such actors and stakeholders as legislators, employees, employers, ordinary courts and high courts of appeal. The approach hinges on the notion of "hot topic" in order to track in probabilistic terms the evolution of the process. Topics corresponding to laws and decisions about the laws are plotted against a timeline, in a way that is similar to the monitoring of the evolution of ideas in a discipline pioneered in Hall et al. (2008). A topic in the timeline heats up when it gets densely populated, namely there are many documents that come into existence in the given interval of time and are associated with the topic, and/or when many other topics point to it through links that can be generated as probabilistic transitions. This information is exploited as an indication that the process has undergone a critical phase as far as the specific topic is concerned, moving from one stage to the next: for instance, if the topic originated from a directive of the European Union heats up in consequence of content corresponding to deliberations of the Italian high court of appeal (Court of Cassation), this is taken as an indication that the given directive may have finished its "run-in" phase in Italy, since the Court of Cassation has often the role of fixing the interpretation of legislation.

4.2 Policy analysis

Policy analysis is yet another domain where the approach to network generation described here has been applied. The term is meant here in the sense of analysis *of* policy, namely of explanation of policies and of the results deriving from their application. (Policy analysis in the sense for analysis *for* policy provides the complementary activity of formulating candidate policies and of articulating the arguments to their support.) At the heart of the methodology there is the idea of comparing the "simple world" of the idealized expectations of the policy maker, with respect to a policy that s/he has issued, with the "complex world" deriving from its implementation. Consequently, results can be compared with expectations, and the effectiveness of a given policy can be judged with respect to the objectives it wanted to achieve.

In the case study and methodology presented in Pareschi et al. (2014), the concept of "outcome network" is introduced as a tool to measure the effects that were expected from a policy-making with respect to the effects that were actually achieved through its implementation. From a given policy a corresponding outcome network can be mapped out through the knowledge of domain experts. That same knowledge can then be used to classify the effects of implementation of the policy under consideration. For instance, in the case of policies that correspond to laws, the contents of sample documents, issued by deliberating bodies like courts with jurisdiction in the matter, are associated with nodes in the network. It is then possible to go one step further, and extract meaningful and appropriate statistics for the objectives and purposes of policy analysis, by using Probabilistic Topic Models

to populate the nodes in the network and thus treating them as if they were topics. Yet another step leads to the generation of full-fledged networks, with contents whose relevance can be weighted on the basis of their capability to attract links from other contents. In this way, when enough data and evidence have been gathered, it is possible to express on empirical grounds an evaluation about the effective capability of the policy in question to achieve the goals originally set by the policy maker. If the topics corresponding to the outcome nodes of the network are populated according to proportions that differ significantly from the expectations of the policy maker, then clearly there is need of correcting some aspects of the given policy. By contrast, we clearly have a success if the proportions are met. In the case of policies, like laws, expressed as documents to which related documents can refer, we have further evidence of success if the policy itself becomes a hub node in the wider document network encompassing decisions related to the issues that it addresses.

5 Conclusion

Viewing complex systems as networks offers a promising and formally well-defined framework for the study of complexity, as witnessed by the outburst of interest in the field of complex networks. However, for such an opportunity to become effective, the premises for the analysis of complex networks must be set on firm and objective grounds, and must be able to take advantage of the wealth of empirical evidence deriving from the big data made available in the digital age, and in particular from data that come in textual form. We have shown how the world of text data and of complex networks can be effectively bridged though the use of machine learning methodologies rooted in information theory and information retrieval, and in particular in the techniques derived from the field of Probabilistic Topic Models. Beside the practical implications that this coupling offers, we have shown how it provides also a satisfactory and plausible completion of the state-of-the-art of complex networks by extending the possibility of observing and analyzing the conditions of emergence of complexity to the basic building blocks of networks, namely to their nodes and links.

References

Bavota G, Gethers M, Oliveto R, Poshyvanyk D, De Lucia A (2014a) Improving software modularization via automated analysis of latent topics and dependencies. ACM Trans Softw Eng Methodol. doi:10.1145/2559935

Bavota G, Oliveto R, Gethers M, Poshyvanyk D, De Lucia A (2014b) Methodbook: recommending move method refactorings via relational topic models. IEEE Trans Softw Eng 40(7):671–694. doi:10.1109/TSE.2013.60

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84. doi:10.1145/2133806.2133826

Chang J, Blei D (2009) Relational topic models for document networks. Ann Appl Stat 4(1):124–150. doi:10.1214/09-AOAS309

Gethers M, Poshyvanyk D (2010) Using relational topic models to capture coupling among classes in object-oriented software systems. In: Proceedings of international conference on software maintenance (ICSM), Timisoara

Granovetter M (1983) The strength of weak ties: a network theory revisited. Sociol Theory 1:201–233. doi:10.2307/202051

Hall D, Jurafsky D, Manning CD (2008) Studying the history of ideas using topic models. In: Proceedings of the conference on empirical methods in natural language processing, ACL 2008, pp 363–371

Harel D, Koren Y (2001) On clustering using random walks. In: Proceedings of foundations of software technology and theoretical computer science, Bangalore, India, Dec 13–15. doi:10.1007/3-540-45294-x_3

Kawaguchi S, Garg P, Matsuhita M, Inoue K (2006) MUDABlue: an automatic categorization system for open source repositories. J Syst Softw 79(7):939–953. doi:10.1016/j.jss.2005.06.044

Kuhn A, Ducasse S, Girba T (2007) Semantic clustering: identifying topics in source code. J Inf Softw Technol 49(3):240–243. doi:10.1016/j.infsof.2006.10.017

Li M, Vitányi P (2008) An introduction to Kolmogorov complexity and its applications, 3rd edn. Springer, New York. doi:10.1007/978-0-387-49820-1

Linstead E, Lopes CV, Baldi P (2008) An application of latent Dirichlet allocation to analyzing software evolution. In: Proceedings of the 2008 7th international conference on machine learning and applications, IEEE Computer Society 2008, pp 813–818

Newman M, Girvan M (2004) Finding and evaluating community structure in net works. Phys Rev E. doi:10.1103/PhysRevE.69.026113

Newman M, Barabási AL, Watts D (2006) The structure and dynamics of networks. Princeton University Press, Princeton

Panichella A, Dit B, Oliveto R, Di Penta M, Poshyvanyk D, De Lucia A (2013) How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms. In: Proceedings of the 35th IEEE/ACM international conference on software engineering (ICSE'13), Vancouver

Pareschi R, Rossetti M, Stella F (2014a) Tracking hot topics for the monitoring of open-world processes. SIMPDA. doi:10.13140/2.1.2910.1761

Pareschi R, Toffoletto F, Zica P (2014b) Outcome networks for policy analysis. NAiL. doi:10.13140/2.1.1667.2320

Rama GM, Santonu S, Heafield K (2008) Mining business topics in source code using latent dirichlet allocation. In: Proceedings of the 1st India software engineering conference (ISEC '08). doi:10.1145/1342211.1342234

Rossetti M, Pareschi R, Stella F, Arcelli Fontana F (2014) Integrating concepts and knowledge in large content network. New Gener Comput 32(3–4):309–330. doi:10.1007/s00354-014-0407-4

Rosvall M, Bergstrom C (2008) Maps of random walks on complex networks reveal community structure. In: Proceedings of the National Academy of Sciences of the United States of America, pp 1118–1123. doi:10.1073/pnas.0706851105

Steyvers M, Griffiths T (2007) Probabilistic topic models. Handbook of latent semantic analysis, pp 424–440

Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to grow a mind: statistics, structure, and abstraction. Science 331(6022):1279–1285. doi:10.1126/science.1192788

Thomas S, Adams B, Hassan A, Blostein D (2010) Validating the use of topic models for software evolution. In: Proceedings of conference on source code analysis and manipulation (SCAM). doi:10.1109/SCAM.2010.13

Thomas S, Adams B, Hassan A, Blostein D (2014) Studying software evolution using topic models. Sci Comput Program 80:457–479. doi:10.1016/j.scico.2012.08.003

Tian K, Revelle M, Poshyvanyk D (2009) Using latent dirichlet allocation for automatic categorization of software. In: Proceedings of conference on mining software repository (MSR), Vancouver. doi:10.1109/MSR.2009.5069496