Social-Collaborative Inductive Reference Model Mining in a Knowledge-Based Organization

Andreas Sonntag[1]

[1] Saarland University, 66123 Saarbrücken, Germany

Abstract. The aim of reference model mining is to support the efficient execution of process instances. Social network analysis has a great potential for reference model mining as it reveals social and functional relations, which are critical for the efficiency of collaborative business processes. This study demonstrates and evaluates an approach to applying social network analysis to a human aspect of reference model mining. The approach is based on a dynamic performer network, which is an evolving social collaboration network in a knowledge-based organization. For this purpose, agent-based simulation is applied to a longitudinal dataset concerning researcher collaboration in an internationally renowned 'center of excellence' for industry-oriented research in the field of artificial intelligence. The resulting performer network can be used as a reference model for efficient researcher collaboration, and it is reusable for future process execution of similar organizations.

Keywords: Reference Model Mining, Human Aspects of BPM, Researcher Collaboration.

CEUR-WS.org/Vol-3115/paper3.pdf

1 Introduction

One important requirement for individuals who participate in a multitude of business processes is the availability of business process models which can be executed efficiently. A reference model should represent an efficient and reusable implementation of processes in an organization, simplifying internal structures to reduce the complexity and resources needed for business process mining [1], [2]. Our research community ignores the influence of social collaboration between people working on processes, called “performers” in the following. Since business processes entail social processes, it becomes essential to employ social network analysis for reference model mining in the knowledge work domain. Previous research on collaboration around knowledge-based processes in the field of business process management are the following: Tomasello and colleagues [3], [4] investigate the formation and performance of collaboration networks by evaluating over time the link formation events involving a knowledge flow between the collaborating parties. [5] introduce a method to interpret the workflow using social networks inferred by interviews and questionnaires with employees. The authors seek to identify gaps between the management view on processes and their actual execution. [6] introduce a concept for the derivation of a reference process model from knowledge-based

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

process models. Thereby, a performer network with topological success structures, holding for a corpus of given process models denotes a reference process model. [7] introduce the concept of "generated performer networks" for optimizing the efficiency of process models to determine "topological success properties" applying agent-based simulation on a given event-driven process chain.

Originated on the static performer network concept of [6], we introduce a procedure to derive a reference model from a process execution observed in a knowledge-based organization. This reference model consists of a reference social collaboration topology that should be similar to a real organization, independent of an organization-specific performer assignment to process functions. The reference model is able to explain the evolvement of the past organization and to show its re-usability by predicting the organization's future evolvement concerning social collaboration topology and efficiency. The paper is structured as follows: First, we describe fundamental terms and methods, then we explain our approach; this is followed by the evaluation of our concept with the experimental design, its evaluation and the discussion of our findings. Subsequently, a conclusion brings this work to a close.

2 Performer Networks

Social network analysis (SNA) investigates similar social structures in organizations, especially in the communication behavior between individuals, applying mainly methods of graph analysis [8]. SNA should be applied on aggregated data when increasing the observation scales [9]. In addition, structural network properties are to the fore towards the outcome of the actual human relationships [11]. Hence, the social topology of people has, traditionally, a greater impact on the result of their collaboration than certain aspects of their collaboration such as e-mailing frequency, sympathy or locations. Many SNA theories and studies are based on simplified but plausible and broadly examined social structures, particularly the formation of clusters/groups, the emergence of hierarchy, sparsity and short paths [9], [12]–[15].

A process model (short PM) is represented as an event-driven process chain. This is a directed graph structure, consisting of a set of edges that indicates an order between a set of process functions and operators. Each process function has a label, which is an expression in natural language and describes the action(s) to be done at this point in the process flow. Events and other (meta) information that might be provided is not considered. A performer network (short PN) extends a PM by a social network of collaborating performers. Performers are agents working on process functions with a set of capabilities in a PN. Capabilities for the purposes of this study are simplified as a set of mappings between a process function and a number indicating the extent of being capable/efficient to work at this function. Every mapping in [0;1] is possible, even mappings with nonexistent functions in order to represent ineffective performer/capability combinations. PNs are formalized as social networks which contain performers as nodes having a unique ID, social connections, capabilities and the ability to work at a process function to which he is assigned to; and two kinds of edges: social edges that connect performers and functional edges that connect performers with process functions (*assigned_nodes*). The other arrows

are process edges, connecting process nodes. Additional definitions: A path is a sequence of edges that connects two nodes. The degree of a node is the number of directly connected neighbors. The mean degree of a PN is the averaged degree over all performers. The average clustering coefficient is the actual number of edges between an agent's neighbors divided by its possible number, averaged over all performers. A PN's density is defined as $\frac{2m}{(n(n-1))}$ with *n* the number of performers and *m* the number of social edges in the PN.

[16] see process efficiency as time and money saving potential for the process execution. [16] require, all processes to follow a sequence of tasks over time timepoints. From a resource-oriented perspective, efficiency can be seen as the minimal use of resources for a certain goal. A comparable definition for the efficiency of an organization is summarized by [17], defining the term "organizational effectiveness" as more than a financial profit but also as an efficient work of employees and managers for the outcome of the organization. Social collaboration is a resource in the scope of organizational structure. In the context of process execution in an organization, a goal is the completion of a sequence of tasks that is based on the process design. The time needed for the goal to be reached depends on the coordination of the actors participating in the effort to complete the tasks. Other resources that cannot (inter)act autonomously such as inanimate goods (steel, paper etc.) can be assumed to have a constant influence on the goal achievement as they can only be used and not contribute an effort. This means that only actors can influence their utilization. Actors in turn need collaboration to utilize resources, especially if different resources require different capabilities to deal with them. As an indicator for the impact of hierarchy on a PN, we measure the number of key performers for each year. A key performer is, adopted from the hub definition by [18], defined as an individual with a degree over three standard deviations above the mean degree. Hierarchy is constituted by a very small minority of key performers, standing against a vast majority of performers with a degree around the mean degree. This structure is often observed in real social networks [13].

A formal and quantitative effort/efficiency definition of a PN is introduced and explained in detail by [7] and [6] as an algorithm based on social network analysis and agent-based simulation: The efficiency of a PN can be computed and optimized only in combination with a PM or a set of PMs. Efficiency in this formal context means how little effort the performers need to undertake to complete one or several tasks simulated simultaneously through a given PM. The effort for the performers to complete the tasks is the sum of capabilities needed by the performers to reach the tasks effort function by function following the process control flow. Neighboring performers help each other with half their common capability. PN effort describes the sum of effort for the whole PN to complete all tasks. The efficiency of a PN in combination with a PM, called "PN efficiency", is defined on the basis of [7] as 1-*PNeffort*/(*max_effort*) with *max_effort* as the maximal possible PN effort referring to a PN consisting of only one incompetent performer, who has a universal capability of 0.1, assigned to all process functions. The PN efficiency definition punishes a PN with many capable performers. A different definition of efficiency, which we outline as activity intensity, comes from [3], [4] as the number of collaboration events (tasks

in our context) on which an agent worked in a time window in ratio to the total number of collaboration events involving all agents in the network in the same time window, averaged over all agents. Finding an efficient PN around PM(s) requires finding an efficient combination of social topology, functional topology and capability distribution. We establish a social topology using the power cluster network generator by [19] which is an extended version of the Barabasi-Albert model [12] that replicates all of the plausible social structures mentionned above. This network generator was tested by [7] to be very effective for generating efficient PNs around a variety of process models.

3 Approach

Our goal is to optimize the efficiency of the social collaboration topology, measured by means of PN efficiency [7], to let a set of performers collaborate to accomplish all pregiven tasks over a period of time. Our approach is based on agent-based simulation [20] and social network analysis [21], [22]. For the subsequent experimental design, we make the following assumptions for the approach, which are also limitations:

1. All performers work 100 percent on all of their tasks. They have a constant minimal competence for other tasks.
2. An organization is reduced to performers, tasks, social- and functional assignments.
3. All tasks follow the same process model; the model topology implies serial and parallel work.
4. All tasks have the same effort of 1.0, simulated or not.
5. An organization grows at a linear pace in terms of number of tasks and performers.
6. The efficiency of social collaboration is measured with activity intensity [4] and PN efficiency [7].
7. Time periods are discrete and equidistant, independent of the exact effort needed.

Algorithm 1. Simulates a Timeline of Performer Networks (*PNT[t : timepoint]*) to Fulfill a Given Set of Tasks

With *simulateTimeline*(*initial_Pnumber*,*tasks*) the evolvement of the social links between the performers, the distribution of their capabilities and process assignments are simulated. *initial_Pnumber* is the number of performers in *initial_timepoint*. *tasks* is a set of tasks where a task is a process instance or, in other words, one complete execution of a process model. As many performers are created as in the observed organization at *initial_timepoint*. The tasks to do here are those that were done by the observed organization at *initial_timepoint (tasks[1988])*. The procedure of simulating the evolvement of social edges is based on the extension of the BA model by [19].

Now, the particular tasks are simulated through the generated PN following the approach of [7]. Therefore nesesary capabilities are distributed over the performers (*distributeCapabilities*). Every performer's capability for a process function is normal distributed with *N(0.5,1)* and constrained co-domain to [0;1]. This gives every performer a chance of 65.5 percent to be more competent than the minimum competence of 10 percent. The assumption here is that every employee has always a minimal knowledge of all process functions. Thus, the flexibility of employees in organizations to substitute other colleagues is taken into account. Next, the performers

are assigned to the process functions (*assignPerformers*). Thus, a performer can pick his process assignment according to his best capability and other assignments with the probability proportional to the extent of the corresponding capability. With *distributeTasks*(*tasks*, *performers*), now the tasks are uniformly distributed over the performers, so that each task is assigned to $\frac{|tasks|}{|performers|}$ performers. Then, *optimizeEfficiency* determines and optimizes/minimizes the PN effort of the PN at that time point with the implementation of [7] and [6]. The optimization reestablishes the social edges by tuning the network generation parameters from [19]. In the next step, the algorithm continues with the next time point (*initial_timepoint + 1*) in which a new performer network is generated.

4 Experimental Design

We implemented our approach into a Java-based research prototype. The hardware configuration for the execution of the evaluation scenario comprises 64 AMD Opteron(TM) 6272 processors @ 1.40GHz and 16GB of RAM. As an evaluation scenario, we employ all scientific publications of the German Research Center for Artificial Intelligence from its foundation with 12 authors in 1988 until 31.12.2017. The organization is the biggest research center in the area of artificial intelligence worldwide, both in terms of number of employees and of external funds; it belongs to Germany's prestigious "Centers of Excellence". According to the most recent data, it has 480 highly qualified researchers and 376 graduate students from more than 60 countries; they are working on 180 research projects. Many of those research projects, which produced a total of 7520 scientific publications, written by a total of 6704 authors, are co-operations with industry.

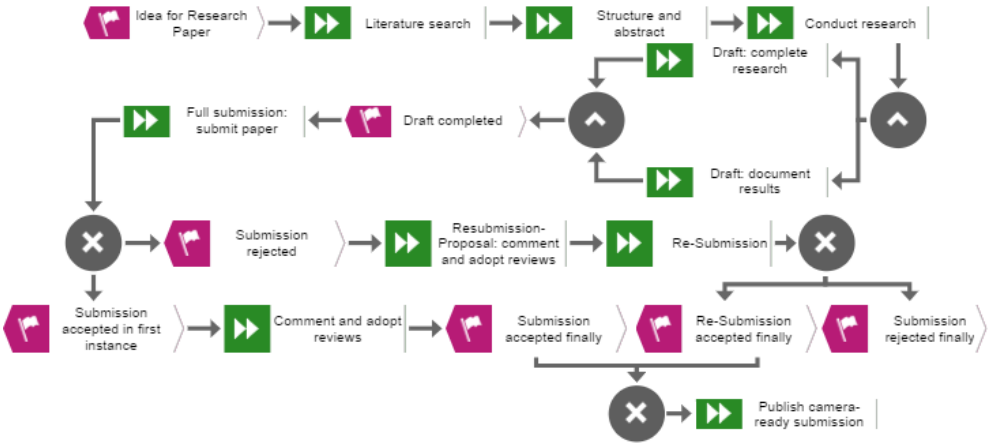

Fig. 1. Process Model for Research Publication

The number of publications rose from 8 in 1988 to 370 in 2017, with an annual average increase of 12.48 (standard deviation: 41.82). The PM in figure 1 describes the minimal requirements for the typical publishing research papers process which are explicitly documented in the organization. Only authors are performers. Each

publication is a task for the performers to fulfill. With *simulateTimeline*(12,*tasks*), for each publication, a PN is derived from the present co-authorship data (not with the social network generator) consisting of the authors involved and the year of publication. The set *tasks* corresponds to all 6904 publications mentionned above. The first PN has 12 performers as there were 12 authors in 1988. For each year, performers are connected if they are co-authors. As a result, we have a timeline of PNs consisting of one PN of all co-authorships for each year. In the following, this timeline is called "observed PNT". *simulateTimeline*(12,*tasks*) is also executed in the same parameterization but with generating the social edges. The result is referred to as reference PNT. To pronounce the design implication again, both PNTs are simulated with *simulateTimeline*, each with the same assumptions.

In order to validate, the results of our experimental evaluation scenario must be reproducible (reliability), must explain the model quality (internal validity) and must produce a result that is generalizable/transferable (ecological validity) [23]. The reliability is reached by test-retest, the repetition of all PN efficiency simulations within the approach. The model quality is quantified by the statistical effect between the starting parameters and the model parameter evolvement. For proofing the generalizability of the approach, the significant correlation between both efficiency measures (see section 2) of the observed PNT and the generated reference PNT has to be shown. We also predict the organization's future evolvement by the same procedure as the reference PNT is developed. For the prediction, the observed PNT from 2010 until 2017 forms the basis for the reference PNT to be evolved from. The prediction is limited to 2022 as our memory capacity is reached at this point of computation effort.

5 Evaluation

The scenario described above yields two simulation results, observed- and the reference PNT, both covering the timespan in which the evolved organization existed. As stated in section 1, we want to explain the evolvement of efficiency in the observed organization with a reference PNT, generated by the simulation described in section 3. Figure 2 compares activity intensity, PN effort and PN efficiency between reference (ref) and observed (obs) PNT over time. Topological properties (average clustering, density and mean degree) between reference and observed PNT are also compared. The average slopes of all efficiency values and the topological properties have, over all years, the same sign. The PN effort for the task execution correlates with the number of tasks over time ($r = 0.97$, $p < 0.01$) for both PNTs. The PN effort increases for both PNTs but is much greater for the observed PNT. In the observed PNT, density and PN efficiency correlate with $r = -0.99$ ($p < 0.01$). This correlation for the reference PNT amounts to $r = -0.91$ ($p < 0.01$). The correlation between activity intensity and PN efficiency in the observed PNT amounts to $r = 0.49$ ($p < 0.01$). The same correlation for the reference PNT amounts to $r = 0.35$ ($p < 0.05$). The independence of the PN efficiency from individual time windows has to be ensured because we want to show that the efficiency of performer collaboration depends on the social collaboration structure and not on the arrangement of single time windows.

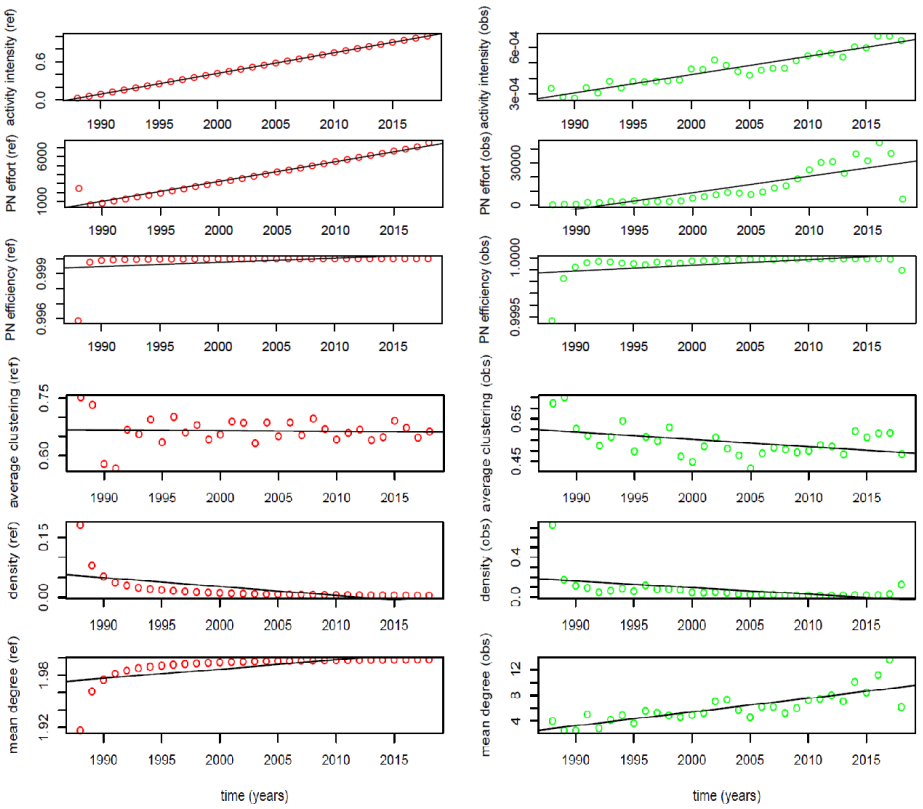

Fig. 2. Reference (ref) vs Observed (obs) Performer Network Timeline with Fit Line

Therefore, we tested hypothesis that the mean of all PN efficiency values for each time point is significantly different to the rolling mean over all possible time windows of a 3-years width. The hypothesis was rejected with $p < 0.05$. The number of key performers evolve almost linearly over time in the observed and the reference PNT. On average per year, the number of key performers grows by 0.21 (standard deviation: 9.6) in the observed PNT. In the reference PNT, 0.1 key performers supervene as a mean (standard deviation: 1.74). The more tasks are to be done, the more key performers appear in both PNTs. For both PNTs, the following variables have an impact on the number of key performers (in descending order): tasks, PN effort, activity intensity. All influences are strongly positive $r > 0.7$ ($p < 0.05$). The PN efficiency has no significant influence of $r = 0.25$ ($p < 0.1$) on the number of key performers for the observed and the reference PNT. This insignificance is caused by the PN efficiency to suddenly and over-linearly increasing with more than 2 key performers.

We predict the organization's future evolvement from 2010 until 2022 and compare it to the observed evolvement in the same time window. The observed PN efficiency stays almost constant at 0.9999 between 2010 and 2017. Our predicted PNT becomes 0.0003 percent more efficient over time. The activity intensity increases linearly over time for the observed and the predicted PNTs. The absolute

values however are different. In our prediction, the activity intensity is always greater than 0.93, whereas the observed activity intensity lies between 0.0003 and 0.0007. The PN effort decreases from 6384 to 4905 in our prediction. In the observed PNT, the PN effort ranges from 25096 to 36756. The average clustering coefficient in the predicted PNT is on average 123 percent higher than in the observed PNT between 2010 and 2017. Density is falling from 0.0061 to 0.0057 in the predicted PNT and increasing from 0.02 to 0.03 in the observed PNT. The average mean degree increases by 0.002 percent for the predicted PNT and 33 percent for the observed PNT.

6 Discussion

Activity intensity and effort increase almost linearly for the observed and reference PNT which means that the reference PNT reaches, for each point in time, an equal or even higher efficiency reached by a lower effort than the observed PNT. All topology parameters, as they are plotted in figure 2, indicate the same slope over time for the observed and reference PNT. The mean degree for the observed PNT increases linearly over time while its reference counterpart reaches its maximum already after 10 years. For both PNTs, the density decreases over time and correlates with the increasing PN efficiency. The density in the observed PNT is much greater than in the reference PNT for all time points while the density within the observed clusters is much higher than in the reference clusters. The reference PNs have sparse clusters, which are but densely inter-connected. Because of the significant correlation between density and PN efficiency, the density of inter-connection between clusters drives the collaboration efficiency, more than the intra-connectivity of teams. The significant correlation between activity intensity and PN efficiency speaks for the generalizability of our approach as both efficiency measures indicate an increase of effective collaboration effort. That means, processes in a knowledge-based organization can be modelled efficiently based on the connection of social and functional reference topologies found by this approach.We predict the organization's future evolvement until 2022. The positive evolvement of activity intensity and PN efficiency over time indicates the evolvement of an efficient social topology around the publication process based on the observed PN in 2017. During the same period, the PN effort decreases in the predicted PNT, in contrast to the observed PNT.

Our explanation for this contrast is the PNs in the predicted PNT to have much more social edges between clusters than the observed organization has. Meanwhile, in the observed organization, on average, more key performers appear than in our reference PNT. That implies the observed performers reached their tasks with a similar efficiency but with more PN effort and more densely connected key performers. Our reference PNT, including the predicted PNT, reaches a smaller PN effort by more social edges between clusters. That way, less collaboration effort respectively fewer social edges are needed between the performers at a common process function to be efficient. The key performers seem to play an important role in the organization's collaboration coordination. Most key performers are managers, for example research group leaders, which means that the team-overarching cooperation between managers is a critical structure for efficiency. This means, that the social link

generation from our PNT simulation procedure (see algorithm 1) can be used to reproduce an efficient collaboration topology in an evolving knowledge-work organization. Our PN simulation can thus be seen as a reference for the efficient placement of personal around a process to be executed in an evolving organization. Translated into a recommendation for modelling efficient collaboration, a performer network should attract more team members around managers and become less dense over time. This evolvement is supposed to lower the costs for social communication by shorting paths in a growing organization.

Our assumptions for the PN efficiency simulation, the simplification of the co-authorship process and the constrained generalizability of the co-authorship towards knowledge-based business processes in general entail limitations of our PN model and our findings. The social environment of the authors, their resource/knowledge allocation and transfer are not taken into account. In addition, our approach has no explicit time limit for the end of the PNT. This means that the simulated organization can take a longer or a shorter time to complete all given tasks.

7 Conclusion

In this paper, the performer network concept of [7] is applied on a set of tasks executed by real collaborative knowledge workers in order to generate a dynamic performer network that completes the given tasks efficiently. By the comparison of the performer network efficiency to a different measure of collaboration efficiency, the activity intensity [4], topology structures of collaboration, similar to the observed over time, were replicated. We tend to regard the performer networks replicating this topology as generalizable references, which may be used by practitioners as guidelines for inferring efficient performer networks around process models of other knowledge working organizations. Furthermore, our approach can quantify the trade-off between team size vs density vs hierarchy vs efficiency-critical social links for the (re)design of processes in such organizations. In particular, this includes the practical issue of determining the number and position of team/division leaders and knowledge/information transfer hubs necessary for a certain process. Applying our "reference topology", this issue can be resolved before the process is even established. Thus, the risks and costs for the process execution become more transparent and controllable.

In future work, we aim to compare our results to further real-world performer networks by using a larger and more diverse data set such as evaluating event logs of the execution of business processes. In particular, we want to understand how exactly process model topologies affect the performance of the assigned performers.

References

1. vom Brocke, J.: Design principles for reference modelling: Reusing information models by means of aggregation, specialisation, instantiation and analogy. In: Innovations in

Information Systems Modeling: Methods and Best Practices, University of Muenster, Germany (2009).
2. Scheer, AW., Nüttgens, M.: ARIS Architecture and Reference Models for Business Process Management. In: Business Process Management - Models, Techniques, and Empirical Studies, W. van der Aalst, J. Desel, and A. Oberweis (eds), pp. 376–389, Springer, Berlin (2000).
3. Tomasello, MV., Tessone, CJ., Schweitzer, F.: The effect of R&D collaborations on firms' technological positions, Proc. 10th Int. Forum IFKAD, no. June, pp. 260–276 (2015).
4. Tomasello, MV.: Collaboration networks : their formation and evolution, ETH Zurich, (2015).
5. Kim, ED., Busch, P.: Workflow Interpretation via Social Networks, Springer, Cham, pp. 241–250 (2016).
6. Sonntag, A., Fettke, P., Loos, P.: Inductive Reference Modelling Based on Simulated Social Collaboration, (2017).
7. Sonntag, A., Fettke, P.: Efficiency Of Generated Performer Networks In Collaborative Business Process Models. In: 18th IEEE Conference on Business Informatics 1, pp. 26–34 (2016).
8. Adamic, L., Adar, E.: How to search a social network, Soc. Networks, vol. 27, no. 3, pp. 187–203 (2005).
9. Easley, D., Kleinberg, J.: Networks, crowds, and markets: Reasoning about a highly connected world, Cambridge University Press (2010).
10. Watts, DJ., Dodds, PS., Newman, J.: Identity and Search in Social Networks, Science (80)., vol. 296, no. 5571, pp. 1302–1305 (2002).
11. Johnson-Cramer, ME., Parise, S., Cross, RL.: Managing Change through Networks and Values, Calif. Manage. Rev., vol. 49, no. 3, pp. 85–109 (2007).
12. Barabási, AL., Albert, R.: Emergence of scaling in random networks. In: The Structure and Dynamics of Networks, Princeton University Press, pp. 349–352 (2011).
13. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 177–187 (2005).
14. Nadel, SF.: The Theory of Social Structure, vol. 23, no. 2. (1957).
15. Li, M., Fan, Y., Chen, J., Gao, L., Di, Z., Wu, J.: Weighted networks of scientific communication: The measurement and topological role of weight, Phys. A Stat. Mech. its Appl., vol. 350, no. 2–4, pp. 643–656 (2005).
16. McKinty, C., Mottier, A.: Designing efficient BPM applications : a process-based guide for beginners, 1st ed. O'Reilly Media, Inc, USA (2016).
17. Schäfermeyer, M., Grgecic, D., Rosenkranz, C.: Factors influencing business process standardization: A multiple case study (2010).
18. Goldenberg, J., Han, S., Lehmann, DR., Hong, JW.: The role of hubs in the adoption process, J. Mark. a Q. Publ. Am. Mark. Assoc., vol. 73, no. 2, pp. 1–13 (2009).
19. Holme, P., Kim, BJ.: Growing scale-free networks with tunable clustering, Phys. Rev. E, vol. 65, no. 2, p. 26107 (2002).
20. Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems, Proc. Natl. Acad. Sci., vol. 99, no. 3, pp. 7280–7287 (2002).
21. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences, J. Inf. Sci., vol. 28, no. 6, pp. 441–453 (2002).
22. Wasserman, S., Faust, K.: Social network analysis: Methods and applications, vol. 8. Cambridge university press (1994).
23. Campbell, D., Stanley, J.: Experimental and quasi-experimental designs for research. London: Ravenio Books (2015).