Pathways to a drug: A mixed methods analysis of emergence

Arho Suominen*

VTT Technical Research Centre of Finland,
Vuorimiehentie 3, PL 1000
02044 Espoo, Finland
E-mail: arho.suominen@vtt.fi

Ozgur Dedehayir

Queensland University of Technology, School of Management, Level 9, Z Block, 2 George Street, Brisbane, 4000, QLD, Australia
E-mail: ozgur.dedehayir@qut.edu.au

* Corresponding author

Abstract: Despite calls for a broader methodological base, the technology management and forecasting literature remains short of studies synthesising quantitative and qualitative methods. This paper addresses this gap by employing a mixed method approach to study the case of Taxol, a revolutionary drug developed to fight ovarian and breast cancer. The paper shows in detail how a data mining based network model can be synthesised with qualitatively created event narrative to understand the development of a technology and its ecosystem. As a result, the paper shows an npartite network of Taxol drug ecosystem containing over 4000 nodes and 11000 edges. The network enables decision makers to understand the process, interaction of actors and lock-in mechanism in the process of technological development.

Keywords: Mixed methods; Technology analysis, Innovation ecosystem; Taxol

1 Introduction

Approaching technological development from either a technology management and forecasting point of view or the more recent innovation ecosystem (IE) approach, studies have pointed to the lack of research taking advantage of mixed methods. In technological forecasting (TF), this call has been persistent, with scholars such as Martino (1993) looking towards analysis that integrates qualitative narrative with quantitative modelling. Meanwhile, in IE research, effort has gone into visualizing large datasets and focusing on interactions between actors (e.g. Basole & Karla 2011), although implications from the analysis appear to be lacking.

With the rise of the ecosystem view, we have seen a number of studies, particularly data driven, focusing on linkages between actors. For example, Guan & Liu (2017) look at the collaborative knowledge networks using patent based indicators. Being data driven, the study is limited with the proxy of interaction based on metadata and not having the

qualitative narrative to support actual interaction. Scientific non-patent references are also a widely adopted tool to create linkages between technology and research, but even the state-of-the-art limit to around 15 percent of data (patents) showing explicit links (van Raan 2017) and research on data-driven implicit links are only emerging (Ranaei et al. 2017). Studies have also looked into finding areas of technological convergence through network ecology (Lee et al. 2016). However, the data driven approach is limited in this instance to the quality of patent data used being an unknown. A mixed methods approach could alleviate data quality issues by highlighting inventions that have had a significant impact on the technological pathway, not considering each instant of patenting activity similarly. This is of particular relevance in areas with high patenting intensity. Visual analytics and network can enable complexity in the system and help understanding the evolution of a specific technology (Kay et al 2017). Existing, highly data-driven, studies are confronted with limitations set by proxies (e.g. Guan & Liu 2017), coverage (e.g. van Raan 2017) and quality (e.g. Lee et al. 2016). With the increase in our data analytics and visualization capabilities Martino's (1993) call for mixed methods is increasingly timely.

We address the problem of how to integrate a robust qualitative and data-driven quantitative approach to holistically model the events, actors, interactions and developments during the emergence of a technology and its ecosystem. To the authors' knowledge, there have only been a handful of examinations of ecosystem birth although with limited mixed method approaches. For example, Garnsey and Leong (2008), illustrate how new ventures in the pharmaceutical sector negotiate constraints presented by the selecting environment during the process of bringing a medicinal product to the market. The authors utilise the notion of a 'transaction environment' to analyse a firm's interaction with other actors in its immediate vicinity, showing the influence of regulatory institutions in promoting or prohibiting the innovations of pharmaceutical firms, the mediation of customer preferences by physicians and hospitals, and the legitimization of organizational activities by public opinion. In a more recent study, Thomas and Autio (2014) examine the emergence of six digital service platform ecosystems. Their results suggest the centrality of four activities for ecosystem emergence: (i) resource activities - the acquisition and management of resources by the hub firm; (ii) technological activities - the design and provision of technologies; (iii) institutional activities - the establishment and implementation of rules of engagement; and (iv) context activities (e.g. regulatory activities) - stimuli sourced from the environment that influence the operation of the ecosystem. These contributions, however, cannot be classified as mixed methods studies, as they primarily employ a qualitative approach.

For our empirical study, we selected the pharmaceutical context that allowed us to readily identify the birth phase of the drug innovation ecosystem, which we define as the period spanning from the initial discovery of an active compound until its first commercial application as a drug. Our study focused on the case of Taxol, a ground-breaking drug commercialised by Bristol-Myers Squibb for the treatment of breast cancer, and the ecosystem that gradually formed around this innovation from 1967 until 1992. We acquired data from archival sources using both qualitative and quantitative methods to trace and understand the events that took place throughout the development of Taxol and its ecosystem.

2 Background

Emergence is a "self-organizing" process incorporating actors, observable events, and

measurable outputs. Theoretical and managerial interest in understanding the process has been long-standing (as early as Wells 1902). The theoretical framework for analysing emergence stems from TF, which developed through the work of researchers such as Kahn, the "father" of scenario analysis, Dalkey, Helmer and Rescher (e.g. Dalkey, 1967), developers of the Delphi method, and Ayres (1969) on technological forecasting. To date, systematic approaches to analysing the potential of technologies have remained macro-level studies, although management focused literature is emerging (e.g. Rohrbeck & Gemünden 2011).

TF is methodologically driven, but rooted in the discussion on technological change and theories such as the evolutionary theory of technological change (ETTC). ETTC offers a robust, but controversial construct to understand emergence and ecosystems. Drawing on the Darwinian model of evolution, ETTC extends the concept to the complex-system of technological development. It essentially builds a link between TF and IE, which we refer to as heterogeneous networks of organizations that collaborate in producing holistic technological systems (Moore, 1993). ETTC sees technology as an improvement through intelligent means, allowing for intentional and random processes. This process explains the survival of the fittest, "...but it cannot explain the arrival of the fittest" (De Vries, 1904). In this study we turn our attention to the arrival, or rather the emergence of a technology and its ecosystem, through the arising of structures, actors and properties in the self-organizing process.

Moore (1993) suggests that innovation ecosystems progress through four life cycle phases, namely, birth, expansion, leadership, and self-renewal (or death). In our paper we address only the birth phase, which crucially centres on the product and service requirements of the customer, and subsequently the definition of a value proposition that can satisfy these requirements. A prominent role that emerges during this phase is that of the 'ecosystem leader'. An actor that assumes this role undertakes the setting of a shared, grand vision, which aims to secure the cooperation of other organisations to provide complementary products and services essential for the delivery of holistic value to the customer. For instance, the birth of Apple's (and its rival, Tandy's) PC (personal computer) ecosystem in the late 1970s underlined the *de facto* leadership role of hardware companies (Moore, 1993). The importance of a central figure for the livelihood of innovation ecosystems is echoed by other scholars as well. Highlighting the platform-based design of ecosystems, Gawer and Cusumano (2002) pinpoint the governing role platform leaders, such as Microsoft and Intel, have historically played in their respective ecosystems. Similarly, Iansiti and Levien (2004) employ the analogy of a keystone, a vital species in biological ecosystems, to represent the role of regulating the overall function of the innovation ecosystem.

Despite their significance and ability to exert substantial power and command a greater share of overall profits (Moore, 1993), actors which enact the ecosystem leader role represent only a small biomass or population of the ecosystem as a whole (Iansiti and Levien, 2004). Constituting the larger bulk of the ecosystem are actors that assume the generic role of the 'niche player'. As in biological ecosystems (the analogical origin of the term), niche players have specialised functions (or a narrow sphere of expertise), which enable their contribution towards the holistic objectives of the innovation ecosystem. Given the role niche players play in the ecosystem, they may also be referred to as 'complementors', which help the ecosystem leader expand the realms of its application. For example, Intel and Microsoft are platform leaders in the PC ecosystem because they assume great authority in the architectural design of the PC system and subsequently govern a plethora of complementors, which produce complementary, platform-specific hardware and software products (Gawer and Cusumano, 2002).

The literature does not, however, confine ecosystem actors to organisations, acknowledging that business communities are social systems and therefore comprise individuals as well (e.g. Sodhi and Tang, 2008). Moore (1993) illustrates the important role of individuals through visionaries such as Ransom E. Olds who helped establish the automobile ecosystem by the beginning of the 20th century. Bahrami and Evans (1995) similarly pay tribute to the roles played by entrepreneurs in negotiating risks to start-up new ventures, venture capitalists who form alliances in co-investing in promising firms, and scientists who wish to commercialise their inventions.

The emergence of IE is therefore likely to comprise a multitude of actors operating at different levels of analysis, which renders the analysis of this process fairly complex. Nevertheless, by taking an innovation ecosystem perspective, we allow the emergence process to take place without restriction. We subsequently move slightly away from similar (and overlapping) constructs such as national systems of innovation that take national boundaries into consideration (e.g. Lundvall, 1993), regional systems of innovation which focus on innovation activities within regional boundaries (e.g. Cooke et al., 1997), the sectoral systems of innovation that study activities within sectoral boundaries (e.g. Malerba, 2002), and technological systems that focus on the generation, diffusion and utilization of technologies by networks of agents (e.g. Carlsson & Stankiewitz, 1995). In turn, TF approaches emergence through ex post analysis, and using historical data, explains the emergence of complex systems and future pathways. Qualitative methods in turn allow the highlighting of events in the emergence process, although grasping large complex systems through a qualitative narrative is challenging. Studies have thus called for mixed approaches in quantifying emergence, specifically using data-driven approaches to quantify events and interactions. This study therefore focuses on two goals:

- How mixed methods can be used to create a system model of technological emergence
- How the model can be employed to a case, to show birth and development of an ecosystem of actors, events and technological pathway.

3 Method and Data

For our empirical study, we selected the pharmaceutical context that allowed us to readily identify the birth phase of the drug innovation ecosystem. Following the contribution of Garnsey and Leong (2008), we defined this phase as the period spanning from the initial discovery of the drug (or its basic chemistry) until its first commercial application. Our study focuses on the case of Taxol, one of the best-selling anti-cancer drug of all time, commercialised by Bristol-Myers Squibb for the treatment of breast cancer. The development of Taxol and its IE took place from 1967 until 1992, coinciding with three sequential trial phases that were mandatory to administer prior to the drug's availability for patients, in other words commercial application. Our analysis of this timeframe is a mixed qualitative and quantitative analysis synthesised to a network representation of the complex system of an emergent technology and its ecosystem.

The quantitative data is gathered from the EuropePMC database, containing life sciences article, patents and clinical guidelines, roughly 31.9 million records. The database was accessed programmatically and text-mined using the Application Programming Interface (API). The API was used to search for documents with terms "Taxol" (the drug's commercial name) or "paclitaxel" (Taxol's chemical name). The search returned 26475 documents, 1264 of which related to the period of Taxol's

development.

The data was manipulated to extract persons, organizations and semantic information (abstracts). Abstracts, were analysed using Latent Dirichlet Allocation (LDA) (refer to Yau et al., 2014, and Suominen, 2015), a machine-learning model that classifies documents based on their content to topics. Abstracts were pre-processed to control for bigrams, exclude punctuations, plurals and stopwords. Short abstracts, fewer than 100 characters, were removed. Pre-processed data were input to the LDA classifier. Persons, organizations and document classes were used to create an npartite network, where nodes are persons, organizations, and topics, while edges denote co-authorships, affiliations, joint publications and links to a topic. This npartite interaction network forms the basis of complex system mapping. The process was implemented using the Python programming language.

For the qualitative component of our study, we acquired data from archival sources to trace and understand the events that took place developing Taxol. To identify relevant documents we employed a search strategy firstly using the Google search engine, using the terms "Taxol" and "paclitaxel". We filtered through the results collecting the most pertinent documents for analysis, which included a wide array of sources such as books, corporate announcements, news, publications, videos, and websites. We complemented this cohort of documents by conducting a second search in three reputable, online databases (ISI Web of Knowledge, Elsevier's Scopus, and Google Scholar). After scouring through the abstracts of this collection, we selected a number of documents that the authors deemed as being most pertinent for a more in-depth study. Our selection protocol particularly aimed at acquiring documents with a broad range of publication years, and also those authored by individuals central to the development of Taxol.

We commenced our analysis by developing a chronology of events that took place during the emergence of the Taxol ecosystem from the textual data (sourced from archival documentation). For each event, we noted the date, actor involved, and the actions of the actor. We then coded the information embedded in this chronology and placed "conceptual labels on the data" through our own understanding of the text (Corbin and Strauss, 2008). This qualitative data was aggregated to the quantitative network, adding nodes for persons, organizations, and events. Edges were updated to include links to events and co-mentions. The resulting dynamic network graph was used as a system model of Taxol development, visualized using Gephi software.

4 Results

Overview of Taxol and its ecosystem

The story of Taxol – the commercialised, pharmaceutical drug – begins with a series of activities that led to the initial discovery of a chemical compound possessing the potential to combat cancerous cells in the human body. These precursors included the US government's initiation of a screening program in 1960 to identify antitumor agents in plants. Two years later, in 1962, Arthur Barclay collected samples from a variety of plants in the Pacific Northwest on behalf of the USDA (US Department of Agriculture), which contained samples of the Pacific Yew (Taxus brevifolia) tree, the source of Taxol's chemical compound. Barclay's samples arrived at the laboratories of Dr. Monroe Wall at the USDA in 1964, who demonstrated the cytotoxicity of the chemical agent (paclitaxel) acquired from the bark of the tree. This result encouraged the initiation of a more purposeful Taxus brevifolia collection and research effort, leading to the isolation of the

chemical's purified fraction in 1966. The first mention of Taxol appears then a year later, in 1967, in a report to the American Chemical Society. We consider this as a milestone in the search and discovery stage of Taxol's development, and the triggering event that commenced the evolution of the Taxol innovation ecosystem, even though the discovery of Taxol as an active compound was disseminated to the scientific community by Wani, Wall, Coggen, and McPhail four years later in 1971.

In studying Taxol's emergence we limited our focus to developments that took place between the early 1960's to the early 1990's. During this timeframe our main data source EuropePMC does not contain any patents, only science publications. Figure 1 shows the dynamics of the publications, which were qualitatively sliced to three periods of interest: early growth until 1983, plateau from 1984 to 1990 and rapid growth in 1991 and 1992. The early growth stage contains approximately 11 % of the publications, the plateau phase 59 % and the rapid growth phase 30 % of publications.

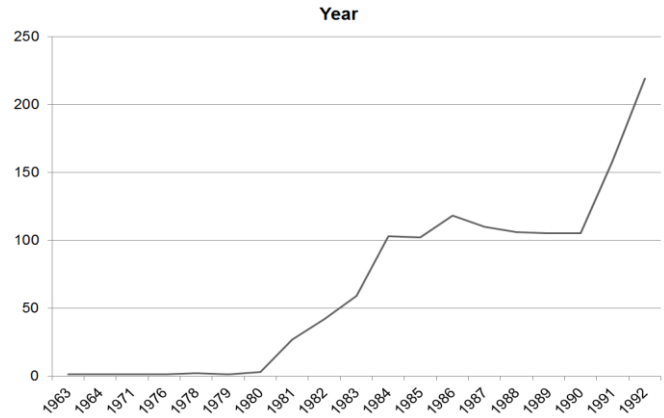

Figure 1 Count of publications during the emergence period. The data consists solely of publications. Source: Europe PMC, Calculations: authors.

The publications which had an abstract, in total 1104, were analysed using the LDA algorithm. The objective of the analysis was to classify the documents based on the latent topics presented in the text. Like for example K-means analysis, LDA requires the researcher to set the number of topics prior to the analysis. To estimate a suitable number of topics, we analysed the KL Divergence, a relative entropy measure, for different number of topics. In practical terms, this analysis looks for a number of topics where the KL Divergence measure would be low, in comparison to other number of topics. Seen in Figure 2, the KL Divergence value has a sharp drop near four and 14 topics. The data was analysed with both 14 and four topics. As the topic concentrated heavily on a few words when the analysis was run with 14 topics, four topics was selected as the final number of topics.

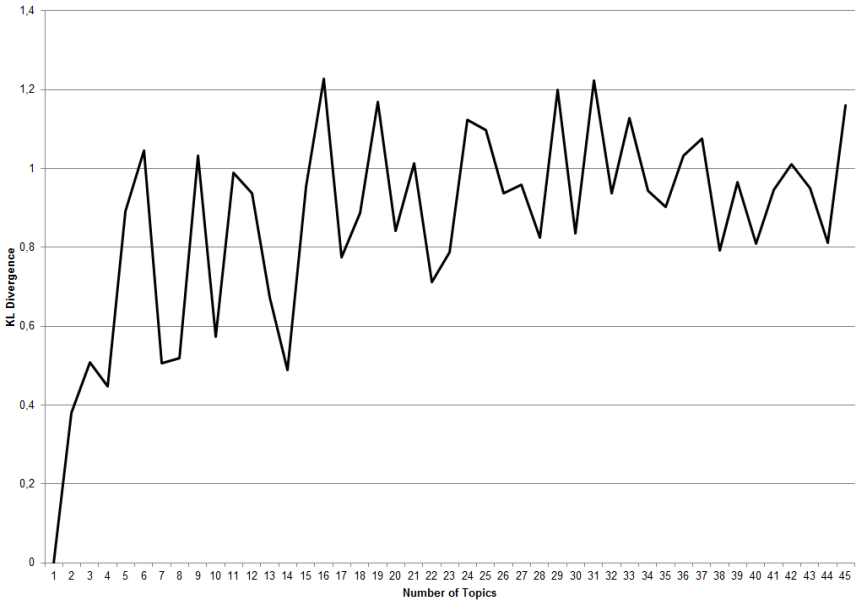

Figure 2 KL Divergence values at different topic values. Topics visualized from 1 to 45, horizontal axis, and KL Divergence value in the vertical axis. Number of documents in the analysis is 1104. Calculations: authors.

The LDA algorithm clustered the documents to four relatively equally sized communities. Topic 1 includes 260 documents, Topic 2 199 documents, Topic 3 295 documents and Topic 4 350 documents. Figure 3 shows the classified topic growth yearly. Topic 1 shows a sharp decrease being the lowest percentage share topic. The topic includes publications focusing on in vivo and in vitro testing of Taxol. Topic 2 is defined by Taxol's impact on the microtubule, a part of cytoskeleton cell component. Both Topic 3 and Topic 4 focus on microtubule-associated proteins (MAPs), which is the most probable word in both topics. Topic 4 has no mention of clinical trials. Topic 3, at a title level, focuses on Phase I and Phase II testing of Taxol. Looking at Figure 3, the topic focusing on clinical trials of Taxol increases by the end of the time series.

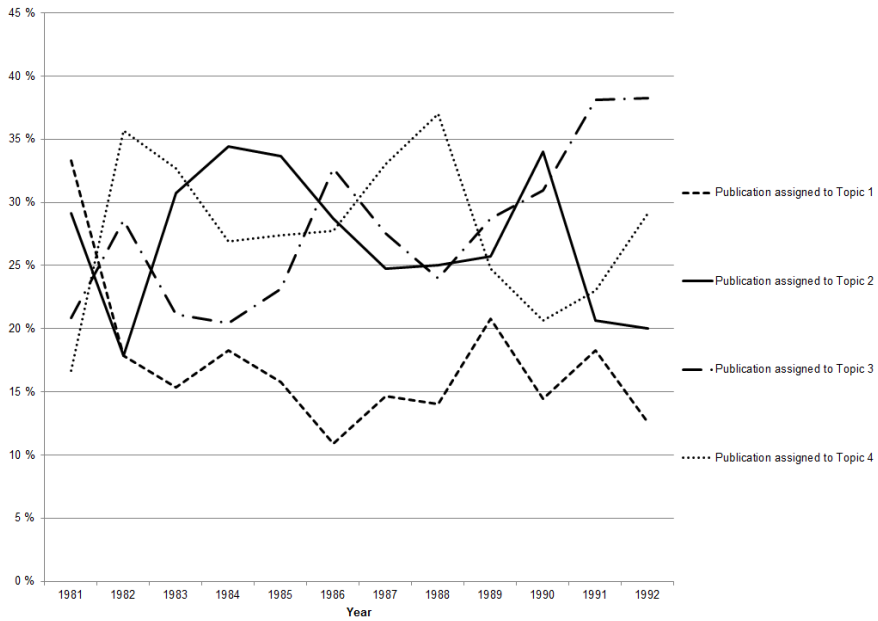

Figure 3 Hard classified document classification as a time series. The values are percentage share of total publications per year Calculations: authors.

From the emergence period we identified 458 separate events, majority of which are from the late period of our time series. For these events we were able to identify 53 individual actors (i.e. persons) and 93 organizational actors. Both person and organizational actors also contain a class "multiple/ambiguous", which was used when unable to identify the individual or organization taking part in a specific event. The count of events identified can be seen in Figure 4.

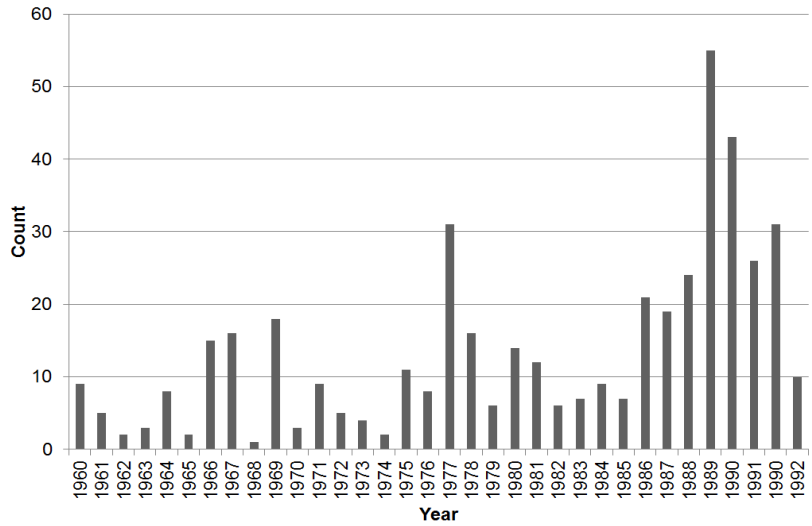

Figure 4 Histogram of identified event counts by years of the emergence process. Calculations: authors.

The events and documents were converted to an npartite network, where nodes are researchers, persons, publications assigned to topics 1 through 4, events and publications not assigned to a topic. Nodes are connected by edges based on publication affiliation, authors and person/researchers connected to events. In total the network has 4698 nodes connected by 10160 edges[1]. The average weighted degree of the network created is 7.367, its diameter 27, density 0.001 and average path length 8.919. The weighted degree measure gives us an understanding on the number of links between the nodes, weighted by the strength of that connection. In the network, the weighted degree being as high as it is results from a few nodes in the graph being highly connected, whereas the majority have only a few connections. Diameter, which describes the largest number of edges which must be travelled in order to go from one node to another, gives a good indication that the network is extremely sparse, as a relatively high amount of nodes need to be passed through to link the farthest nodes. This argument is also supported by the low density and high average path length in the graph.

Focusing on a qualitative assessment of the graph, it is clear that there are isolated subnetworks that are strongly interconnected in the event and publication space. Seen in Figure 5 is one tightly formed cluster from the network. This graph highlights researcher (lilac) and person in (light gray), events (red), organization's (green) and publications assigned to Topic 3 (blue) and Topic 4 (dark grey). Through a qualitative evaluation we can establish that the dense intersection of nodes in the figure relate to the early stage testing of Taxol as a drug.

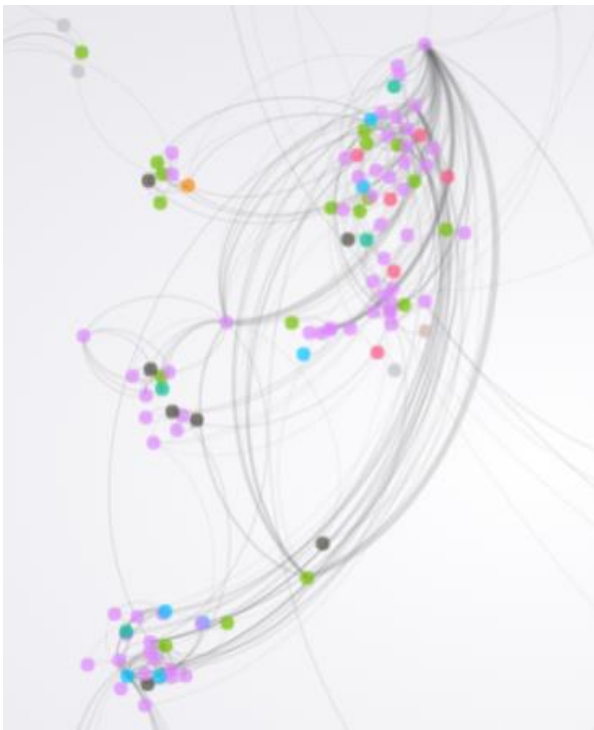

Figure 5 Excerpt of the whole network graph, which shows a dense subnetwork focusing on early stage drug development.

[1] The full graph is available at https://tinyurl.com/yb6fv2dh

Interestingly the qualitative analysis of the Taxol network shows numerous small subnetworks, where a clear leader is challenging if not impossible to identify. An excellent example in the dataset is event 4588, where the identified organizations striving to synthesize Taxol seem isolated from each other rather than forming a knowledge sharing community of researchers.

Synthesis of Taxol ecosystem emergence

The synthesis of the quantitative and qualitative study of the Taxol case underscores the roadblocks to technology and ecosystem emergence, which must be sequentially overcome to guarantee progress. During its very nascent period the ecosystem was extremely small, comprising only a few actors (e.g. Monroe Wall and his colleagues, and the USDA). At the same time, a lot of uncertainty surrounded the future of the ecosystem. A number of concerns emerged, which would collectively cause the NCI (National Cancer Institute) to lose interest in the isolated compound, and prevent Taxol from entering trials until 1977. The first hurdle was encountered by Wall and his colleagues, who had begun work on understanding the structure of the compound in 1967. It soon emerged that the concentration of the active compound in the raw material (i.e. the bark of the Taxus brevifolia tree) was low, and that the supply of the tree itself was limited. There also appeared to be difficulties in the extraction and isolation of the agent given its low solubility. Together, these problems cast doubts into the minds of the stakeholders as to future endeavours considering the drug's large-scale production. In parallel, paclitaxel was observed as having only a limited range of impact on the standard cancer models implemented by the NCI. Interest in Taxol additionally waned because of its average performance, and the falsely perceived familiar action of Taxol's mechanism, in comparison to other chemicals that were being assessed at that time. It was not until the NCI introduced additional cancer models in 1975 that paclitaxel could demonstrate its true effectiveness. Two years later, in 1977, paclitaxel was accepted as a candidate for preclinical development.

These compounding barriers to ecosystem development were essentially overcome by two individuals. First, Dr. Monroe Wall continued to push for the potential of the isolated compound, and his insistence eventually allowed him to prove the activity of the chemical agent against a test case (B16 melanoma). The second individual was Dr. Susan Horwitz, who, together with her colleagues in 1979, proved the unique biological effects and mechanism of the chemical agent against cancerous cells. The actions of Wall and Horwitz served to reinvigorate the compound's potential, retriggering the development of the nascent ecosystem, the fragility of which had been fuelled by the inherent scepticism of key stakeholders such as the NCI, who made vital decisions on whether to continue or kill such projects. Finally, in 1983, approval was given for the commencement of clinical trials, which concurrently shifted the ecosystem onto a new and different set of activities.

The removal of the NCI roadblock resulted in a surge in scientific work in the early 1980s, supported by our quantitative data, which nevertheless plateaued in 1985. Our qualitative examination revealed this loss of momentum was caused by concerns of material supply during drug trials. Collectively, the quantitative and qualitative data confirmed the supply problem delaying development approximately five years.

The clinical phase I and II trials were conducted between 1983 and 1986. Nevertheless, this period was also marked by problems that threatened to end ecosystem development. These included the poor stability, hypersensitivity, and toxicity of the drug, which caused safety issues and the abandonment of many trials. The supply problem also

reappeared between 1987 and 1989, when the success of the drug against ovarian cancer created a buzz within the scientific community and the general public, increasing demand and causing a concern as to the availability of stock and raw materials. The media, which not only pronounced Taxol as a miracle drug, but also emphasised the lack of raw materials for the production of the drug. The Pacific Yew tree, to source of paclitaxel, was indeed the central issue, given its rarity and slow rate of its growth. The media had inevitably and importantly sparked dialogue between relevant actors (e.g. doctors, politicians, environmentalists), who coalesced about the controversy, and triggered further scientific search to find alternative solutions.

The NCI, now assuming a central, ecosystem leader role in the further development of the Taxol ecosystem, targeted the problem of supply by issuing a CRADA (Cooperative Research and Development Award) to motivate competition for higher quantities of supply. In 1991, the award was given to Bristol-Myers Squibb, a large pharmaceutical firm interested in the drug's potential. This was a watershed event in the emergence process as our quantitative data indicates. After several years of monotonous development, a second surge in research was initiated as the community's doubts about the drug's future were put to rest. In parallel, the NCI tackled the supply issue by organising workshops in 1990 and 1992, to generate additional research on sourcing the raw material. Altogether, the emergence of these problems invited different organisations to enter the milieu as complementors by offering specialised functions (e.g. Hauser Chemical Research to collect bark under contract), and the ecosystem began to grow in terms of actors and their connections. Meanwhile, Bristol-Myers Squibb filed for and received an NDA in 1992. We observe this as an important transition point in the birth of the ecosystem, as the leadership begins a hand-over from the NCI, previously driving research and development as well as problem resolutions, to Bristol-Myers Squibb, a large pharmaceutical company with expertise and interest in drug commercialisation. At the same time, the year 1992 brings the birth phase of the Taxol ecosystem to a close.

The supply controversy emerged as the pivotal matter in the overall birth of the Taxol ecosystem. While the NCI in alliance with Bristol-Myers Squibb followed a path dependent search protocol, alternative paths were being created by others. In 1981, Potier and Greene in France had started a program to produce Taxol from the needles of the Taxus baccata tree, a close relative of the Taxus brevifolia tree, through a semi-synthesis process. Finally in 1988, these scientists published the successful results of their process, which provided an avenue of resolving the ecosystem's supply problem. Yet another avenue would open up from the total synthesis of Taxol. In 1993, with the anticipation of receiving FDA approval for the drug, there has been a notable intensification of a race to synthesise Taxol. A year later, in 1994, Holton and Nicolaou, two scientists independent of each other, announced the total synthesis of Taxol. These scientists had essentially sown the seeds of another solution to the supply problem, and simultaneously began the redirection of the ecosystem's developmental path, beyond the birth phase.

5 Discussion and Conclusions

The main contribution of our paper is methodological in its nature. Responding to scholars such as Devezas (2005) our complex system approach offers an evolutionary vantage point to understand emergence and ecosystem creation, far extending the workhorse of TF, the logistic equation. This methodological advancement has theoretical implications to ETTC. Our analysis provides a systemic view of emergence as an evolutionary process, emphasising the role of links between actors, events, and

metamorphosis in the emergence process.

Understanding the emergence of technologies and their ecosystems is crucial for organizations investing in their initial development and construction. Such understanding can help organizations anticipate and resolve challenges threatening to dissolve the volatile processes and nascent networks, while aiding others to make decisions whether to join the ecosystem. From a policy perspective, a systemic understanding of the process of emergence and ecosystems, and the role of governance, can assist decision making concerning policy interventions. With any given ecosystem, such governance may be enacted by institutional actors at the national (e.g. Lundvall, 1993) and/or regional (e.g. Cooke et al., 1997) levels, as well as seminal actors (e.g. the ecosystem leader).

The study has several limitations. In making an effort to draw insight from multiple datasets and both qualitative and quantitative analysis we struggle with integrating the results to one holistic view. In the case at hand, the process involved data drawn from an API and mixed sources of qualitative analysis results. Making an effort to find common authors, organizations and events and to understand the intensity of links between the nodes, we relied on automated methods (e.g. string matching). These methods are not perfect, and produce false positives and missing true connections. We made efforts to correct these by evaluating random data points in the graph. However, errors are likely to exist even after our efforts. Another limitation is the extent of analysis done for the events. A more thorough analysis focusing on identifying corresponding actors to those emerging from the data analysis could yield a stronger network representation. To offset this limitation, we went through the qualitative analysis data a second time, focusing specifically on identifying missing actors behind events.

Notwithstanding these limitations, the method used in the paper showcases an approach to combine qualitative and quantitative methods of analysing technological pathways. We believe that future work can focus on increasing the robustness of this approach and validating its capability to create insights through other case studies.

Acknowledgements

Tekes—the Finnish Funding Agency for Innovation supported this work with the research grant awarded for the project “Integrating Platform Competences toward Network Effects”, as well as the Academy of Finland research grant (288609) “Modeling Science and Technology Systems Through Massive Data Collections”

References

Ayres, R. U. (1969). *Technological forecasting and long-range planning*. McGraw-Hill New York, USA.

Bahrami, H. and Evans, S. (1995). Flexible re-cycling and high-technology entrepreneurship. *California Management Review*, 37(3): 62-89.

Basole, R. C. and Karla, J. (2011). On the evolution of mobile platform ecosystem structure and strategy. *Business & Information Systems Engineering*, 3(5), p. 313

Carlsson, B. and Stankiewitz, R. (1995). On the nature, function and composition of technological systems. In: Carlsson B. (Ed.), *Technological Systems and Economic Performance*. Kluwer Academic Publishers, Dordrecht.

Cooke, P., Urange, M. G., and Extebarria, E. (1997). Regional innovation systems: Institutional and organizational dimensions. *Research Policy*, 4/5, p. 475–493.

Corbin, J and Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory (3rd ed.).* SAGE Publications, Los Angeles, USA.

Dalkey, N. (1967). *Delphi.* RAND Corporation, USA.

Devezas, T.C. (2005). Evolutionary theory of technological change: state-of-the-art and new approaches. *Technological forecasting and social change*, 72(9), pp. 1137-1152.

De Vries, H. (1904). *Species and varieties: their origin by mutation*. The Open Court Publishing Company, Chicago, USA.

Garnsey, E and Leong Y.Y. (2008). Combining resource-based and evolutionary theory to explain the genesis of bio-networks. *Industry and Innovation*, 15(6), pp. 669-686.

Gawer, A and Cusumano, M.A. (2002). *Platform Leadership*. USA: Harvard Business School Press.

Guan, J., & Liu, N. (2016). Exploitative and exploratory innovations in knowledge network and collaboration network: A patent analysis in the technological field of nano-energy. Research policy, 45(1), 97-112.

Iansiti, M and Levien, R. (2004). Strategy as ecology. *Harvard Business Review*, March, pp. 1-10.

Kay, L., Porter, A. L., Youtie, J., Newman, N., & Ràfols, I. (2017). Visual Analysis of Patent Data Through Global Maps and Overlays. In Current Challenges in Patent Information Retrieval (pp. 281-295). Springer Berlin Heidelberg.

Lee, W. J., Lee, W. K., & Sohn, S. Y. (2016). Patent Network Analysis and Quadratic Assignment Procedures to Identify the Convergence of Robot Technologies. PloS one, 11(10), e0165091.

Lundvall, B. A. (1993). *National Systems of Innovation.* Frances Pinter, London.

Malerba, F. (2002). Sectoral systems of innovation and production. *Research Policy*, 31, p. 247–264.

Martino, J. (1993). *Technological Forecasting for Decision Making 3rd ed*. McGraw-Hill Inc

Moore, J.F. (1993). Predators and prey: A new ecology of competition. Harvard Business Review May-June, pp. 75-86.

Ranaei, S., Suominen, A., & Dedehayir, O. (2017, June). A topic model analysis of science and technology linkages: A case study in pharmaceutical industry. In Technology & Engineering Management Conference (TEMSCON), 2017 IEEE (pp. 49-54). IEEE.

Rohrbeck, R. and Gemünden, H.G. (2011). Corporate foresight: Its three roles in enhancing the innovation capacity of a firm. *Technological Forecasting and Social Change*, 78(2), pp. 231-243.

Sodhi, M.S. and Tang, C.S. (2008). The OR/MS ecosystem: Strengths, weaknesses, opportunities, and threats. *Operations Research*, 56(2), pp. 267-277.

Suominen, A. and Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology,* 67(10), pp. 2464-2476.

Thomas, L.D.W. and Autio, E. (2014). The processes of ecosystem emergence. Working paper.

van Raan, A. F. (2017). Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research. Journal of Data and Information Science, 2(1), 13-50.

Wells, H.G. (1902). *Anticipations of the Reaction of Mechanical and Scientific Progress Upon Human Life and Thought*. Chapman & Hall.

Yau, C.K., Porter, A., Newman, N., and Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3), pp. 767-786.