Characterizing Communities of Practice in Emerging Science and Technology Fields

Olga Babko-Malaya, Daniel Hunter, Gregory Amis
BAE Systems
Burlington, Massachusetts USA
{olga.bako-malaya, daniel.hunter,
gregory.amis}@baesystems.com

Patrick Thomas
1790 Analytics
Haddonfield, New Jersey USA
pthomas@1790analytics.com

Adam Meyer
Computer Science Department
New York University
New York, New York USA
meyer@cs.nyu.edu

James Pustejovsky, Marc Verhagen
Computer Science Department
Brandeis University
Waltham, Massachusetts USA
{jamesp, marc}@cs.brandeis.edu

Emerging fields in science and technology are of great interest to innovation researchers, but such fields are often difficult to identify and characterize. This paper outlines a system for identifying a key element of emerging fields: their community of practice, consisting of active scientists and researchers. The system does not simply count these human actors and the interactions between them. Rather, guided by actant network theory, it also examines other non-human actors with which they interact, such as organizations, publications and terminologies. Using quantitative indicators inspired by actant network theory, and derived from features extracted from the full text and metadata of scientific publications and patents, the system attempts to identify communities of practice associated with emerging fields in science and technology. This paper outlines details of these features and indicators, describes how these indicators are combined using Bayesian models, and reports the results of applying these indicators to document sets associated with emerging scientific and technological fields. The results reported in this paper show that system outputs generally agree with subject matter expert judgments with respect to determining the existence of communities of practice, and appear to offer interesting insights into the development of emerging fields.

Keywords—actant network theory, emergence, community of practice

I. INTRODUCTION

In the study of innovation, a great deal of attention is paid to emerging technologies. Such technologies are held to be highly generative [1], with the potential to open up whole new areas of technology and science. This potential – in simple terms, the promise of being the next biotechnology or nanotechnology – inevitably draws interest from various organizations. These include government agencies looking to fund promising new ideas, corporations hoping to gain a foothold in a rapidly emerging field, and investment institutions seeking returns from early investments in key innovators. Emerging technologies have also been a focus of academic research ever since Schumpeter coined the term 'creative destruction' to describe the emergence of new technologies [2], which spawned new industries while destroying old ones. More recently, Christiansen & Bower used the term 'disruptive technology' to describe a new development that disrupts the status quo in an existing technology [3].

Yet, despite this widespread interest in emerging technologies, identifying such technologies remains problematic. The problems are both theoretical and practical. The theoretical issue is how to recognize an emerging technology, without a clear definition of what constitutes such a technology. As noted by Goldstein [4], there is a lack of precision in the meaning of emergence, and even greater ambiguity about how it occurs. The practical issues result from the sheer scale of information available, especially in the electronic age. Researchers and analysts searching for interesting new technologies face the unenviable task of locating meaningful signals among this mass of information. Their task is not aided by the fact that the number of truly emergent technologies is dwarfed by the number of mature, mundane, or failed technologies.

II. SYSTEM OVERVIEW

In an effort to address both the theoretical and practical issues associated with locating and characterizing emerging technologies, we are developing an automated system that

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. Approved for Public Release; Distribution Unlimited.

processes very large collections of scientific publications and patents, extracts features from the full text and metadata of these publications and patents, and constructs quantitative indicators based upon these extracted features.

The indicators employed in the system are designed specifically to locate and characterize emerging scientific and technological fields. The theoretical foundation for these indicators is provided by actant network theory [5]. This theory provides a vision of science and technology as constituted by networks of heterogeneous elements, interconnected by disparate relationships. These networks do not just contain individuals, but also institutions, instruments, practices, terminology, materials, funders, meetings, government organizations, laws, journals, patents, publications, and so on. The membership of elements within such a network, and the nature and extent of the relationships between these elements, is dynamic and constantly changing.

In the idiom of actant network theory, our system's task is to identify, characterize, and evaluate over time the actant networks that comprise emerging technologies and emerging fields. More specifically, this task is to use indicators from the metadata and full-text of publications and patents associated with these fields in order to identify, characterize, and evaluate over time the actant networks of science and technology.

In this paper, we present a model of one aspect of actant networks - the community of practice, which is the networks' human dimension or component. The community of practice is perhaps the key element of the actant networks associated with scientific fields or technologies, since these fields and technologies are not active if no one works in them.

Our model measures five different characteristics of a community of practice that are suggested in previous research as being relevant to emergence [6]. The first three of these characteristics – extent, resilience and connectivity – point to the existence of such a community, and its likely ongoing robustness. That is, for a community of practice to exist, it must be of a certain size (extent) and be connected enough (connectivity) to withstand the removal of key actants (resilience). The other two characteristics – growth and novelty – point specifically to emerging technologies. That is, for a community of practice to be defined as being associated with an emerging field or technology, it must not only exhibit extent, resilience and connectivity, but also growth and novelty. These latter two characteristics differentiate the communities associated with emerging technologies from those associated with mature, stable technologies.

The indicators used to assess the five characteristics listed above go beyond simply the human dimension. They also account for the non-human actants with which researchers interact, such as organizations, publications and terminologies. This follows actant network theory, which posits that human actors do not interact in a vacuum, but rather in the context of other non-human actants. It is thus insufficient to simply count researchers, and the interactions among them, when assessing communities of practice.

This paper contains four further sections. In the first of these sections, we outline the various features extracted from the metadata and full text of scientific publications and patents. We then describe how these features are used to construct indicators designed to assess the five characteristics of communities of practice listed above. Having outlined these indicators, we then demonstrate how they are combined via Bayesian networks to optimize the community of practice model. Finally, we show the results applying this model in practice to sets of scientific publications and patents associated with eight sample technologies, both emerging and non-emerging. These document sets are referred to as Related Document Groups (RDGs).

The results demonstrate the extent to which the outputs of our model concur with the opinions of subject matter experts (SMEs) regarding the presence or absence of communities of practice in the eight sample technologies. The analysis is carried out across six time periods, since there is a temporal aspect to actant networks, and a community of practice in a given technologies may not exist at one point in time, only to develop in a later time period.

III. EXTRACTING METADATA AND FULL TEXT FEATURES

We extract features from the metadata and full text of both scientific papers and patents. The scientific paper collections we currently process include Elsevier (full text articles from 438 journals over 1980-2011, ~4M records), Thomson Reuters' Web-Of-Science® (abstracts of journals and conference proceedings for the same time period, ~40M records) and Pub-Med Central. (full text articles from biomedical journals, ~250k records, dominantly 2008-present). We also processed Lexis-Nexis Patent data which includes granted patents and published patent applications from multiple national patent offices. The features we extract from these sources include:

Feature 1: Documents - this is the most basic feature - i.e. the document itself, whether it is a scientific paper or patent. Scientific papers are disambiguated to avoid mis-counting, while patents are identified by their patent number, which is a unique identifier. Our disambiguation component is based on a combination of blocking and clustering algorithms, where blocking generates an index for each entity, based on its associated data, and groups each entity with the same index into a block and each block is further processed by using a clustering algorithm.

Feature 2: Individuals - we extract the names of authors on scientific papers, and inventors on patents. Names of individuals are disambiguated using blocking functions that exploit other data fields, including organization affiliation, co-authors or co-inventors, subject of research etc - to reduce problems due to common names. Also extracted are the geographical locations of individuals.

Feature 3: Organizations - we extract and disambiguate the names of the organizations with which researchers are affiliated. These are extracted from author affiliations in the case of scientific papers, and assignees (i.e. owners) in the case of patents. Also extracted are the geographical locations of these organizations. In addition, we divide organizations into three types - companies, academic, and government/non-profit (plus individuals in the case of patents).

Feature 4: Citation Links and Indicators - we construct a citation graph based on the citation links between patents,

between papers, and between the two document types. In addition, we derive patent citation indicators that measure the impact, generality and originality of patents [7]. These indicators are normalized to take into account both the age and technology focus of patents. We also determine whether patents qualify for Emerging Clusters. These are clusters of recent patents that are grouped around earlier, highly cited, 'hot' patents. The method for generating these clusters is described in detail in [8].

Feature 5: *Funding Organizations* - papers typically acknowledge funding organizations in their 'Acknowledgments' or 'Funding' section, or in a footnote. Meanwhile, patents sometimes acknowledge funding by government agencies in the 'Government Interest' field. We use regular expressions, keywords and named entity (NE) recognition to identify such funding organizations in both patents and papers.

Feature 6: Abbreviation - Abbreviation relations are used by authors to define short equivalents of longer, often multiword terms. Abbreviations sometimes appear in a single table with a name such as 'List of Abbreviations', and can be read from this table. Abbreviations also appear in the main text, typically in parentheses, following the sequence of words they represent. In such cases, we search backwards from a term preceded by a left parenthesis (the abbreviation) to identify sets of words (the antecedent) that the term is likely to abbreviate. We exploiting the regular relations between abbreviations and their antecedents, for example the first letter in each word of a string, followed by an abbreviation with these first letters (as in Genome Wide Association Study and GWAS).

Feature 7: Terminology - We identify instances of terminology by means of several different methodologies. For example, the antecedents in the Abbreviate relation tend to be either terminology or organization names, and we use named-entity recognition to identify and remove the organizations. We identify noun sequences that are more frequent in documents within the given RDG than they are in general, by comparing noun groups in RDGs with noun groups in some background set, using statistical techniques similar to those reported in [9].

Feature 8: Exemplify - Exemplify relations are 'is-a' relations, as in 'X is a Y'. First, we identify instances of terminology. Then we find instances of these terms in text when they fit templates along the lines of [10]. For example the string 'X, such as Y', suggests that X and Y are in an Exemplify relation, where Y is an example of X.

Feature 9: Opinion: Positive & Significant - authors occasionally indicate that they have a positive view of cited documents, theories, methods or products. To detect instances of such positive views, we used a maximum-entropy-based machine learning approach. Annotators recorded sentiment as well as lexical signals (e.g., "useful") that signaled positive sentiment. The system learned and made predictions based on feature types including: n-grams, occurrence of lexical signals, pronouns, negation (counter-indicators), words like *formula* or *table* indicating illustrative material, parts of speech sequences and dependencies paths linking citations to signals. A similar approach is also used to identify Opinion: Significant relations, in which authors indicate that they regard particular previous research as significant.

Feature 10: Related Work: Contrast & Corroboration - Related Work relations indicate that two concepts are related in some way, for example they are in contrast or they corroborate each other. Our system attempts to find instances where one or both of the contrasted or corroborated concepts are doubled or instantiated by document citations. Our system finds discourse adverbials, conjunctions, verbs, etc. that imply contrasting or corroborative relations between their arguments and assumes that citations contained in their arguments are also being contrasted or corroborated.

Feature 11: Common Topics - we expect an established community of practice to share consistent sets of terminology. To identify sets of terms that are used together and can be leveraged to identify robust terminology use within a community, we have developed a fast, scalable implementation of the online Latent Dirichlet Allocation algorithm (oLDA) originally posed by [13]. This indicator measures the degree to which a community contains topics that are specific and, at the same time, shared across documents produced by the community. Specific means that relatively few documents in the entire corpus discuss this topic. Shared means that many documents from the community discuss the topic. In addition to this indicator measuring distributions across documents, our system adds indicators measuring topic distributions across authors, funders, and organizations.

Feature 12: Document Genre – we interpret genre as the term for any category of scientific literature characterized by a particular style and form. Genre is a cross-domain and cross-topic index, and allows us to look at kinds of texts from diverse topics or fields, regardless of their content. We created an ontology of genres that generalizes over the genres found in the data sets used. This ontology includes 21 types that are arranged in a shallow hierarchy. Some frequent genres are research Article, Review Article, Report, Book Review, Product Review, Commentary, Abstract, Letter, Correction, Case Report, Editorial, Short Communication, and Discussion. Orthogonal to this classification, we use dimensions of analysis like basic type (descriptive, narrative, expository), brow (popular, middle, high), function (inform, persuade, debate), source (academic, industry, legal) and channel (newspaper, journal, report, proceedings, webpage). In particular, we focus on the debate function feature, which reflects the prevalence of debate in a community. Static snapshots of the distribution of genres at a particular moment in time (or time slice) are hypothesized to reflect characteristics of the community of practice. For example, a high debate score and a high relative share of review type genres may indicate more substantial internal discussions in the community. Similarly, changes in the community of practice are hypothesized to co-occur with changes in distribution over genres in the scientific literature generated by the community.

IV. Network Characteristics & Indicator Patterns

Having extracted the features outlined above, the next step is to employ these features to construct indicators related to different characteristics of the community of practice. The rationale for these indicators has been discussed in previous work by the authors (see [6]). We list first the indicators for the three characteristics associated with the existence of a community of practice – i.e. extent, resilience and connectivity.

We then list the indicators associated with the additional characteristics of communities of practice in emerging technologies – i.e. growth and novelty.

A. Extent of a Community of Practice

This characteristic (or 'pattern' in the idiom of the model) measures the size of the actant network. The indicators in this pattern are all counts of actant network elements including researchers, organizations, patents, and characteristic terminology, as well as the extent of high impact actants.

The indicators for Extent of Community of Practice are:

1) Extent of community members

This is the number of individuals who are listed as either (i) an author on a scientific paper, or (ii) an inventor on a patent, in a given RDG (Related Document Group) during a given time period. This indicator is based on the premise that, for a community of practice to exist in a given technology, there must be a sufficient number of people active within it.

2) Extent of patent inventors

This is a sub-element of the community members indicator. It measures the number of unique individuals listed as inventors on patents in a given RDG during a given time period.

3) Extent of researchers

This is also a sub-element of the community members indicator. It measures the number of unique individuals who are listed as authors on scientific papers in a given RDG during a given time period.

4) Extent of unique organizations

This is the number of unique organizations listed on publications in a given RDG during a given time period. It is based on the idea that the greater the number and diversity of organizations involved in a field or technology, the greater its assumed robustness.

5) Extent of funders

This is the number of different funding organizations acknowledged in the text of research papers in a given RDG and time period. The greater the number and diversity of funders, the greater the assumed robustness of the technology.

6) Extent of unique countries

This is the number of different countries (based on author affiliations) listed by authors of publications in a given RDG and time period. Greater geographical dispersion of researchers is regarded as a positive signal in terms of the robustness of the community of practice.

7) Extent of publications

This is determined by counting the number of distinct research papers published in a given RDG and time period. The greater the number of papers published, the greater the assumed robustness of the community of practice.

8) Extent of patents

This is determined by counting the number of distinct patents issued in a given RDG and time period. The greater the number of patents issued, the greater the assumed robustness of the community of practice.

9) Extent of High Impact patents

As a community of practice develops around a technology, key patents that form the foundation for the technology are likely to be cited as prior art by the patents of researchers who develop the technology further. The existence of high impact (i.e. highly cited) patents in a given technology and time period suggest there are many such researchers building on the foundation patents. Conversely, a lack of highly cited patents may suggest that there is a less developed and cohesive community of practice, given the lack of a common foundation for research.

10) Density of Significant publications

This indicator computes the density of Opinion: Significant relations in a given RDG and time period. The presence of numerous mentions of important and significant technologies and documents is an indication that an extended field is developing around particular foundations.

11) Extent of review type articles

The extent of review type articles is determined by counting the number of 'review'-type scientific papers in a given RDG and time period. The presence of reviews is an indication that the field has developed to the point where a review is justified, outlining alternative views from an extended community.

B. Resilience of Community of Practice

The resilience of an actant network lies in its ability to maintain its connectivity and continue to function despite the failure or removal of crucial actants (e.g. star researchers or major funders). To measure resilience, we evaluate the number and diversity of central actants and also the number and diversity of the relationships between these actants. The diversity of actants can be measured by heterogeneity - i.e. the number of different types of actants in the relevant network. For example, as noted earlier, a network with several different types of funders – civilian government agencies, military agencies, large corporations, venture capitalists, and private foundations – is regarded as more robust than a network with a single funder type.

Indicators for Resilience of Community of Practice are:

1) Diversity of Funders

Funders (i.e. organizations funding research) are categorized into three groups: Companies, Academic, and Government/Non-profit, and this indicator measures how many different types of agencies fund research in a given RDG and time period. The greater the diversity of funding sources, the greater the assumed resilience of the technology.

2) Diversity of Patent Assignees

Patent assignees (i.e. owners of patents) are categorized into four groups: Companies; Academic; Government/Non-profit; and Individuals. This indicator measures how many different types of patent assignees are represented within the RDG. The greater the diversity of patent assignees, the greater the assumed resilience of the field or technology

3) Diversity of document genres

This indicator measures how many different genres or document types are represented within the RDG. The greater the diversity of different types of documents within the RDG, the greater the assumed resilience of the field or technology.

4) Diversity of Organizations

This indicator measures how many different types of organizations are represented in the affiliation fields of

scientific papers. The greater the diversity of types of organizations publishing papers in the RDG, the greater the assumed resilience of the field or technology.

5) Average Patent Generality

This indicator computes the mean generality score for all patents in an RDG and time period. It measures the extent to which a technology draws the attention of researchers in other disciplines. The greater the diversity of disciplines interested in a technology, the greater its assumed resilience.

6) Extent of patents with high Patent Generality scores

This indicator computes the number of patents with patent generality scores greater than 1.5. The greater the diversity of disciplines interested in a technology, the greater the assumed resilience of the field.

7) Growth in length of independent claims

This indicator computes growth in mean length of independent claims (in terms of number of words) for all patents in a given RDG and time period. Such an increase may show development in a technology since, as the technology develops, the patent landscape becomes more crowded, and inventors have to use longer claims to delineate their invention's specific niche within this landscape.

C. Connectivity of Community of Practice

This pattern measures the number of relationships and the density of interconnections in a network. To analyze connectivity, we evaluate traffic between actants within network, that is, activity in these relationships and interconnections. Our definition of traffic integrates the rate of information flow (e.g. the spread of terminology) with other factors that characterize the dynamic behavior of actants in the networks (e.g. scientists moving from one organization to another or new scientists entering the field).

Indicators for Connectivity of Community of Practice are:

1) Percentage of review type articles

This indicator measures the percentage of 'review'-type scientific papers in a given RDG and time period. The presence of review articles is an indication that a field has generated enough research to justify a summary of different research themes, and the various connections among them.

2) Extent of publications with authors from different organizations

This indicator is derived by computing the number of publications in a given RDG and time period which have more than one organization in the affiliation field. It thus measures the extent of the connectivity between different organizations active in the technology.

3) Density of co-citations

This indicator computes the density of co-citation relations [11] in a given RDG and time period. It thus measures the extent to which researchers are connected through their reliance on similar earlier research.

4) Density of Opinion: Positive relations

This indicator computes the density of Opinion: Positive relations in a given RDG and time period. It measures the amount of positive sentiment relations between citing and cited documents, which indicates the acceptance of ideas and technologies discussed.

5) Density of contrast relations

This indicator computes the density of Contrast relations in a given RDG and time period. It measures the level of debate in the technology, which is an indication of an active, connected community of practice.

6) Density of Abbreviation relations

This indicator computes the density of Abbreviation relations in a given RDG and time period. When abbreviations are used extensively, this suggests an acceptance of terminologies by a closely connected community of practice.

7) Density of Exemplify relations

This indicator computes the density of Exemplify relations in a given RDG and time period. It is an indication of the interconnections between terminologies, and consequently the acceptance of terms by a connected community.

8) SoftMax of highly Accepted Characteristic Terminology

This indicator measures the soft maximum, over all RDG-characteristic terms, of the fraction of documents in a given RDG and time period that use the term at least once. It thus measures the degree to which RDG-specific terminology has gained wide acceptance, which may in turn be indicative of a closely connected community of practice.

9) Extent of acceptance of Characteristic Terminology

This indicator estimates the degree to which characteristic terms are accepted by a substantial fraction of the community, which may be indicative of a close-knit community. Rather than directly counting the use of terms, this indicator also weights terms by their 'characteristicness' score.

10) Common topics

This indicator measures the degree to which a given RDG in a given time period contains topics that are specific and, at the same time, shared across documents in the RDG. Specific means that relatively few documents in the entire corpus discuss this topic; shared means that many documents in the RDG discuss this topic. The presence of specific, shared topics may be a signal of terminological acceptance, and thus a close-knit community.

11) Multi-Author topics

This indicator measures the degree to which an RDG in a given time period contains topics that are specific and, at the same time, shared across authors in the RDG. Specific means that relatively few authors in the entire corpus discuss this topic; shared means that many authors in the RDG discuss this topic. The presence of multi-author topics suggests that different authors are carrying similar research, and are connected through their use similar terms to describe this research.

12) Multi-Funder Topics

This indicator measures the degree to which an RDG in a given time period contains topics that are specific and, at the same time, shared across funders active in the RDG. Specific means that relatively few funders in the entire corpus fund work on this topic; shared means that many funders within the RDG fund work on this topic. The presence of multi-funder topics may show that different funding bodies are supporting similar research, which is connected by similar terminologies.

13) Multi-Organization Topics

This indicator measures the degree to which an RDG in a given time period contains topics that are specific and, at the same time, shared across organizations active in the RDG. Specific means that relatively few organizations in the entire corpus produce work on this topic; shared means that many organizations within the RDG produce work on this topic. The presence of multi-organization topics may show that different organizations are carrying similar research, and connected through their use similar terms to describe this research.

D. Growth of Community of Practice

One of the basic characteristics of communities of practice associated with emerging fields or technologies is that they are likely to be growing in size. This differentiates these communities from those associated with more mature fields, which are likely to be much more stable. In order to identify growing technologies and fields, we thus evaluate increases in the number and diversity of actants, as well as increases in traffic between these actants.

The indicators for Growth of Community of Practice are:

1) Growth of community members

This indicator counts the number of individuals publishing or patenting in a given RDG and time period, and divides it by the number of individuals publishing or patenting in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in numbers of community members.

2) Growth of inventors

This indicator counts the number of inventors patenting in a given RDG and time period, and divides it by the number of inventors patenting in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in numbers of inventors.

3) Growth of researchers

This indicator counts the number of researchers publishing in a given RDG and time period, and divides it by the number of researchers publishing in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in numbers of researchers.

4) Growth of organizations

This indicator counts the number of organizations publishing or patenting in a given RDG and time period, and divides it by the number of organizations publishing or patenting in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in numbers of organizations.

5) Growth of organization types

This indicator counts the number of organizations types (i.e. companies, academic, government/non-profit) publishing or patenting in a given RDG and time period, and divides it by the number of organization types publishing or patenting in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in numbers of organization types.

6) Growth of funding organizations

This indicator counts the number of organizations funding research in a given RDG and time period, and divides it by the number of organizations funding research in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in numbers of funding organizations.

7) Growth of funding organization types

This indicator counts the number of different types of organizations funding research in a given RDG and time period, and divides it by the number of types of organizations funding research in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in number of different types of funding organizations.

8) Growth of patents

This indicator counts the number of patents issued in a given RDG and time period, and divides it by the number of patents issued in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in number of patents issued.

9) Growth of papers

This indicator counts the number of papers published in a given RDG and time period, and divides it by the number of papers published in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in number of papers published.

10) Growth of countries

This indicator counts the number of countries publishing or patenting in a given RDG and time period, and divides it by the number of countries publishing or patenting in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are particularly likely to experience growth in number of active countries.

11) Growth of genres

The growth of genres is determined by counting distinct genres of scientific papers published a given RDG and time period, and dividing it by the number of genres in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are likely to attract different types of actants, who will contribute different genres of documents.

12) Growth of review articles

The growth of review articles is determined by counting the number of review articles published a given RDG and time period, and dividing it by the number of review papers in the same RDG and previous time period. This indicator is based on the premise that novel fields and technologies are likely to have increasing numbers of review articles written about them.

13) Growth of terminology use

The growth of terminology use is determined by counting the number of characteristic terminologies in a given RDG and time period, and dividing it by the number of characteristic terminologies in the same RDG and previous time period. This

indicator is based on the premise that novel fields and technologies are likely to have increasing numbers of characteristic terminologies, unlike more mature technologies that use a more stable set of terminologies.

E. Novelty of Community of Practice

This pattern is indicative of new domains within technology and science that are attracting increased interest and activity, including newly forming and changing technologies. It evaluates the extent to which people and resources are entering a technology or field, in order to distinguish emerging technologies and fields from more mature fields. This pattern also evaluates domains where actors are importing terminology and resources from other existing domains into the selected field or technology.

The indicators for Novelty of Community of Practice are:

1) Extent of emerging cluster patents

This indicator measures the percentage of patents in a given RDG and time period that qualify for inclusion in emerging clusters, relative to the overall percentage of patents in emerging clusters in the same time period. The presence of emerging clusters in an RDG is indicative of a novel technology attracting a high level of attention.

2) Extent of new terminology

This indicator counts the number of RDG-characteristic terminologies that first appears in a given time interval. Appearance of new terminologies may be linked to novelty within the field.

3) Extent of new inventors

This indicator counts the number of inventors that started patenting on the topic of the RDG for the first time in the given time period. The extent of new inventors in a given field or technology is indicative of a novel field or technology.

4) Extent of new authors

This indicator counts the number of authors that started publishing on the topic of the RDG for the first time in the given time period. The extent of new authors in a given field or technology is indicative of a novel field or technology.

5) Extent of new organizations

This indicator counts the number of organizations that publish papers in an RDG for the first time in a given time period. The premise of this indicator is that novel technologies and fields will be particularly likely to have new-entrant organizations.

6) Average Patent Originality

This indicator computes the mean patent originality score for all patents in a given RDG and time period. Novel technologies may be characterized by original ideas and inventions, whereas mature technologies may feature a higher proportion of incremental ideas and inventions.

7) Extent of patents with high Patent Originality score

This indicator computes the number of patents with the originality score higher than 1.5. Novel technologies may be characterized by original ideas and inventions, whereas mature technologies may feature a higher proportion of incremental ideas and inventions.

V. A Model for the Existence of a Community of Practice

As outlined above, there are three indicator patterns directed to the existence of a community of practice in a given scientific field or technology, and a further two indicator patterns directed to detecting communities of practice specifically in emerging technologies. In the case of the existence question, we had available gold standard data in the form of responses from subject matter experts (SMEs) as to whether communities of practice existed in given technologies in particular time periods. There were a total of eight such technologies and six time periods (1981-1985; 1986-1990; 1991-1995; 1996-2000; 2001-2005; 2006-2010). For example, the SMEs might be asked whether a community of practice existed in Genetic Algorithms in 1981-1985, or in Tissue Engineering in 2001-2005.

The outputs of our model were tested against these SME responses to determine the extent to which the outputs matched the gold standard data (although it should be recognized that, as with any human-based response data, there is the possibility of bias or misunderstanding in the responses from the SMEs). The novelty and growth patterns were not included in this stage of the analysis, since we did not have SME responses to the questions of whether communities of practice were either novel or growing. These two indicator patterns were introduced at a later stage of the analysis, as outlined in the next section of this paper.

Before reporting the results from comparing the outputs of our model with the responses from the SMEs, it is first useful to outline details of the model. Our models are based on Bayesian networks, which are probabilistic graphical models. Probabilistic relationships among variables are captured by a directed acyclic graph. In this graph, the nodes are variables together with a specification of the probability distribution for each variable conditional on values for its parent variables (i.e. variables having edges to the given variable). Such a representation allows for a decomposition of the joint probability distribution over the variables that supports efficient computation of probabilities.

Our models are hierarchical. Indicator variables are linked to pattern variables that are in turn directly linked to the question variable. The pattern variables represent abstractions or summaries of related indicator variables. Such a hierarchical structure has a number of advantages. First, it allows the use of many correlated variables without danger of over-counting. For example, counts of publications and counts of researchers will be highly correlated so that treating them as independent pieces of evidence would overstate their influence. If, however, we take these two variables to be manifestations of a more general community extent variable and link this variable to the question variable, then the correlation between the publication and researcher count variables will be properly accounted for. A second advantage of the hierarchical approach is that it enables more accurate modeling when training data is sparse. With sparse data, estimating the correct, or even an approximately correct, probability distribution for each indicator variable conditional

on the question variable is very error prone. With intermediate pattern variables, theoretical considerations can be brought to bear to help shape the distribution in conjunction with whatever ground truth data is available. A third advantage to a hierarchical structure is that explanations for answers are more comprehensible when framed in terms of meaningful patterns above the level of individual indicators. Finally, the intermediate patterns will often be of interest in and of themselves, independently of how the question is answered.

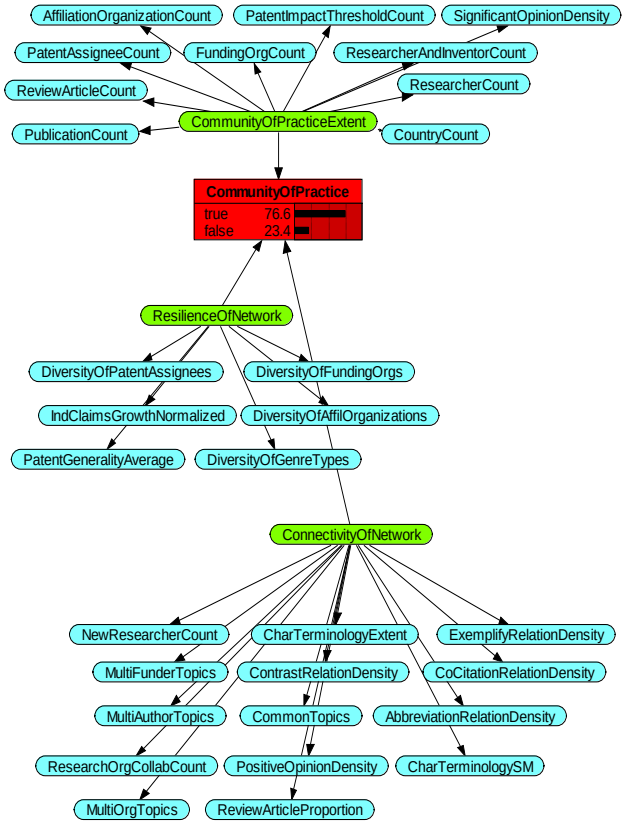

Fig. 1 – Model for Existence of Community of Practice

Figure 1 shows our model for addressing the question 'Was there a community of practice around <concept> during <time period>?'. It should be noted that other indicator patterns may also have informational value regarding this question. However, the additional patterns we considered were not used, either because analysis showed their information value to be low, or because modeling their relationship to the existence of a community of practice question proved difficult and error-prone. The model therefore does not make use of those indicators applicable only to patterns not present in the model. In addition, we performed an evaluation of indicators with potential relevance to a model pattern, eliminating those having insignificant predictive value. The indicators were evaluated on ground truth data by using Monte Carlo methods to estimate the mutual information between an indicator variable and the challenge question answer. The most predictive indicators according to this evaluation include: diversity of genres, extent of organizations, extent of community members, density of abbreviations, density of exemplify, and diversity of funders.

Each subgraph in Figure 1, consisting of a pattern node and its children, is a naïve Bayes model. Updating in a naïve Bayes model is particularly simple. Evidence takes the form of an assignment of a value to a child variable and each piece of evidence contributes independently to the posterior distribution over the parent variable.

The pattern variables are linked to the question variable through a 'noisy and' distribution. This requires some explanation. The classical 'logical and' relation involves a deterministic relationship between a set of Boolean conditions and their 'logical and' conjunction. That is, each of the Boolean conditions is necessary for the truth of the conjunction; failure of any one condition deterministically implies the falsity of the conjunction. The notion of logical conjunction can be generalized, however, to a notion of probabilistic conjunction in which each of the Boolean conditions in the conjunction is only necessary to some degree, where 'necessary to some degree' is cashed out in terms of there being a certain probability that the probabilistic conjunction will fail if the Boolean condition fails (for logical 'and' conjunction, this probability is 1).

For our community of practice existence model, we used the following Boolean conditions over the pattern variables as 'noisy and' conditions:

1. CommunityOfPracticeExtent $\geqslant$ T (where T is an adjustable parameter representing some integer value for the size of the actant network below which we would not count the network as a community; since there can be very small communities, we pegged T at the fairly small value of 30).
2. ResilienceOfNetwork is not low.
3. ConnectivityOfNetwork is not low.

Each of these conditions was considered 'necessary to some degree' for a community of practice to exist. The degrees of necessity were set in such a way that the existence of a community of practice was greater than 0.5 when at least two of the conditions were satisfied, but less than 0.5 if two or more conditions were not satisfied.

Our system was run on 510,000 scientific articles and patents. We first extracted all non-RDG specific metadata and full text features from this dataset, disambiguated people, citations, and organizations, classified organizations, and stored the results in a RDF database. We further processed the RDGs for each of the eight sample technologies for each of the six time periods and extracted RDG specific features, such as terminology and impact. The RDGs included: DNA Microarrays (29K scientific papers and patents); Genetic Algorithms (83K); Cold Fusion (2.1K); Steganography (12K); RF Metamaterials (2.8K); Horizontal Gene Transfer (4.7K); Tissue Engineering (29K); and RNA Interference (66K). We then computed the indicators and ran our model for each of the eight RDGs. Where the model output a value greater than 0.5 for the existence of a community of practice, this was considered a positive answer to the question (i.e. the model states that a community of practice existed in that technology and time period). Conversely, a value lower than 0.5 was

considered a negative output (i.e. the model states that a community of practice did not exist in the given technology and time period).

These outputs were then compared against the responses from the SMEs, and the results are shown in Table 1. There are a total of 47 answers in the 'All' column of Table 1. This represents eight technologies times six time periods, minus one time period/technology for which the data were too sparse for the model to run. The results in Table 1 show that the model worked consistently well with respect to the SME responses. Recall is 0.97 (i.e. 97% of time period/technology pairs with positive SME responses to the question of whether a community of practice existed were also given positive answers by the model). Meanwhile, Precision is 0.79 (i.e. 79% of time period/technology pairs marked with a positive answer by the model were also marked positive by the SMEs); and Accuracy is 0.79 (i.e. in 79% of cases, the responses from the model and the SMEs matched, whether these responses were positive or negative).

TABLE I. COMPARISON WITH SME RESPONSES

	All	Cold Fusion	RF	Tissue Eng
true positives	34	5	3	5
false positives	9	1	1	0
true negatives	3	0	2	1
false negatives	1	0	0	0
recall	0.97	1	1	1
precision	0.79	0.83	0.75	1
accuracy	0.79	0.83	0.83	1

The results in Table 1 suggest that our model shows promise in terms of determining the existence of communities of practice. However, these results are based on a very small data set, largely due to the time-consuming nature of surveying SMEs to generate the gold-standard data. Additional research using more extensive data sets may thus be instructive, in order to determine whether the promising results reported here are repeated for a larger sample of technologies.

Also, as noted above, the existence of a community of practice does not in itself denote an emerging technology, since mature technologies also have communities of practice. Having said this, the model does appear successful in recognizing the transition from absence to existence of communities of practice in given fields and technologies. In itself, this may a positive signal for identifying fields that are emerging, and have communities of practice for the first time.

VI. EXTENDING ANALYSIS TO INCLUDE PATTERNS DIRECTED TO EMERGING TECHNOLOGIES

In addition to tracking the transition from absence to existence of a community of practice, we also examine two specific indicator patterns directed to emerging technologies – novelty and growth. In discussing these patterns, we focus on three of the eight technologies – Cold Fusion; RF Metamaterials; and Tissue Engineering – each of which is generally considered to have been emerging at some point, with Cold Fusion being of particular interest given the controversies surrounding it.

Table 1 shows that in Tissue Engineering, the model output matched the SME response in all six time periods, while in RF Metamaterials and Cold Fusion, the model matched the SME response in five out of six time periods (the model returned one false positive for each of these technologies).Table 2 provides more detail at the indicator pattern level for Tissue Engineering.

TABLE II. INDICATOR PATTERN OUTPUTS IN TISSUE ENGINEERING

	Extent	Resilience	Connectivity	Novelty	Growth
1981-1985	Low	Mod	Mod	FALSE	0-0.25
1986-1990	Low	Mod	High	TRUE	0-0.25
1991-1995	Mod	Mod	High	TRUE	0-0.25
1996-2000	High	Mod	High	TRUE	0-0.25
2001-2005	High	Mod	High	TRUE	0.25-0.5
2006-2010	High	High	High	FALSE	0-0.25

This table reveals that, in the initial time period (1981-1985), there was a very small community of practice (i.e. low Extent). This community then started to grow (see Growth column) and exhibited high Novelty. Growth and Novelty remained high until the most recent time period (2006-2010), when the Novelty pattern became negative. This suggests that, by this point in time, Tissue Engineering had become a relatively stable, mature field, with high Extent, Resilience and Connectivity, but low Novelty. Growth had also started to slow, having been particularly strong in 2001-2005.

Beyond these results at the level of indicator patterns, it is also possible to drill down to the level of individual indicators. For example, we did not identify any funders in Tissue Engineering prior to 1996, but this number rose to 57 in 1996-2000, and increased again to 375 in 2000-2005.. Over the same period, there were also sharp increases in the number of publications, number of patent assignees, and number of Abbreviation relations (as terminologies became more widely accepted).

Similar results to those in Tissue Engineering can be seen with regard to RF Metamaterials, as outlined in Table 3.

TABLE III. INDICATOR PATTERN OUTPUTS IN RF METAMATERIALS

	Extent	Resilience	Connectivity	Novelty	Growth
1981-1985	Low	Mod	Mod	FALSE	0-0.25
1986-1990	Low	Mod	Mod	TRUE	0-0.25
1991-1995	Low	Mod	High	TRUE	0-0.25
1996-2000	Mod	Mod	High	TRUE	0-0.25
2001-2005	High	Mod	High	FALSE	0.25-0.5
2006-2010	High	High	High	FALSE	0-0.25

Again, this field had a small community of practice in the early time periods, but one that started to grow and exhibit strong Novelty. The Novelty pattern again pre-dates the transition from a small to a large community of practice, as it did in Tissue Engineering. As such, this raises the possibility of the Novelty pattern being a leading indicator of fields that are about to emerge.

As in the case of Tissue Engineering, it is possible to drill down to the indicator level to see a similar growth in funders

in RF Metamaterials. We did not identify any funders in RF Metamaterials in 1996-2000, but this number rose to 7 in 2001-2005 and 48 in 2006-2010. This coincided with 1000% and 200% increases in the size of the RF Metamaterials community of practice in 2001-2005 and 2006-2010 respectively. There were also sharp increases in the number of publications, number of patent assignees, and number of Abbreviation relations during this period.

The third technology we highlight here is Cold Fusion, the results for which are shown in Table 4.

TABLE IV. INDICATOR PATTERN OUTPUTS IN COLD FUSION

	Extent	Resilience	Connectivity	Novelty	Growth
1981 - 1985	Low	Mod	Mod	FALSE	0-0.25
1986 - 1990	Low	Mod	High	TRUE	0-0.25
1991 - 1995	High	Mod	High	FALSE	0-0.25
1996 - 2000	Low	Mod	High	TRUE	0-0.25
2001 - 2005	Mod	Mod	High	FALSE	0-0.25
2006 - 2010	Mod	High	High	TRUE	0-0.25

The Cold Fusion example is interesting, because this technology does not follow the 'traditional' linear pattern of non-existence, followed by emergence, followed by maturity. Rather, Cold Fusion emerged very rapidly, was then largely discredited, before re-emerging in a less ambitious form as low-energy nuclear reactions (LENR). This progression can be traced through the indicator patterns in Table 4. Novelty was high in the late 1980s, which is when the famous Fleischmann and Pons paper was published [12]. The Extent pattern then peaked in the early 1990s, before dropping sharply, and recovering to moderate levels in recent years as LENR has developed. As in the other case studies, the Novelty indicator appears early in the development of the community of practice, again suggesting that this pattern may be an early signal of emergence (even if, in the case of Cold Fusion, this emergence was less typical than in the other cases).

Taken together, the results of this section suggest that the Novelty and Growth indicator patterns provide interesting additional information, over and above that provided by the extent, resilience and connectivity patterns. In particular, they provide insights into the dynamic nature of scientific fields and technologies, particularly those that are emerging.

The tables in this section also provide a level of detail on communities of practice that would be very difficult and time-consuming for a human researcher to replicate. For example, it would be difficult for an individual to provide definitive answers to all five questions addressed in these tables – i.e. how big was the community of practice?; how connected and resilient was it?; how novel was it?; how fast was it growing? – especially for multiple time periods. As such, the model described here may be a useful tool for examining communities of practice, and how they evolve over time.

VII. CONCLUSIONS

This paper outlines a system directed to the identification of communities of practice in scientific fields and technologies, particularly those that are emerging. This system uses indicator patterns - based on features extracted from the metadata and full text of scientific papers and patents - to assess different characteristics of communities of practice. The results reported in this paper suggest that the proposed model shows promise in terms of locating communities of practice, based on comparisons with responses from subject matter experts. It also appears to be successful in recognizing the transition from absence to existence of communities of practice in scientific fields and technologies, which is a possible signal for an emerging area of research. In addition to this signal, we also examine two indicator patterns specifically directed to emerging technologies – novelty and growth. These two patterns appear to offer interesting insights into the development of emerging fields, closely tracking the transition from absence to existence of a community of practice. As such, they may be additional useful indicators of emergence. They also add to the output from the model, which as a result offers a level of detail on different dimensions of communities of practice that would be very difficult and time-consuming for human researchers to replicate.

REFERENCES

[1] S. Cozzens, S. et al., "Emerging Technologies: Quantitative Identification and Measurement," Technology Analysis and Strategic Management vol. 22, pp.361-376, 2010.

[2] J. Schumpeter, Theorie der Wirtschaftlichen entwicklung, Leipzing: Duncker & Humboldt, 1912.

[3] C.M. Christensen and J.L. Bower, "Customer power, strategic investment, and the failure of leading firms," Strategic Management Journal, vol. 17(3), pp. 197-218, 1996.

[4] J. Goldstein, "Emergence as a Construct: History and Issues," Emergence: Complexity and Organization, vol. 1, pp. 49-72, 1999.

[5] B. Latour, Reassembling the social: An introduction to actor-network theory, Oxford, UK: Oxford University Press, 2005

[6] D. Brock, O. Babko-Malaya, J. Pustejovsky, P. Thomas, S. Stromsten and F. Barlos, "Applied Actant-Network Theory: Toward the Automated Detection of Technoscientific Emergence from Full-text Publications and Patents," AAAI Fall Symposium on Social Networks and Social Contagion, Arlington, VA, Nov 2-4, 2012

[7] B. Hall, A. Jaffe and M. Trajtenberg, "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools", CEPR Discussion Paper No. 3094, December 2001.

[8] P. Thomas and A. Breitzman, "A Method for Identifying Hot Patents and Linking them to Government-funded Scientific Research", Research Evaluation, vol. 15(2), pp. 145-152, 2006.

[9] R. Navigli and P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites", Computational Linguistics, vol. 30(2), pp. 151-179, 2004

[10] M. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992.

[11] H. Small, "Co-citation in the scientific literature: a new measure of the relationship between two documents," J. Amer. Soc. Inform. Sci, vol. 24, pp. 265-9, 1973.

[12] M. Fleischmann and S. Pons, "Electrochemically Induced Nuclear Fusion of Deuterium," Journal of Electroanalytical Chemistry, vol. 261(2), pp. 301-308, 1989.

[13] M. Hoffman, D. Blei, and F. Bach. "Online learning for latent Dirichlet allocation", Neural Information Processing Systems, 2011