Leveraging Social Information Systems: Using Blogs to Inform Technology Strategy Decisions

Satwiksai Seshasai

Working Paper CISL# 2011-02

June 2011

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E62-422
Massachusetts Institute of Technology
Cambridge, MA 02142

LEVERAGING SOCIAL INFORMATION SYSTEMS: USING BLOGS TO INFORM TECHNOLOGY STRATEGY DECISIONS

by

Satwiksai Seshasai

S.B., Computer Science and Engineering – Massachusetts Institute of Technology, 2001
M.Eng., Electrical Eng.and Computer Science – Massachusetts Institute of Technology, 2002
S.M., Technology and Public Policy – Massachusetts Institute of Technology, 2005

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in ENGINEERING SYSTEMS
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2011

Signature of Author.....
Engineering Systems Division
August 2010

Certified by .....
Stuart Madnick, John Norris Maguire Professor of Information Technology, MIT Sloan School of Management and Professor of Engineering Systems, MIT School of Engineering
Thesis Supervisor

Certified by .....
Joseph Sussman, JR East Professor of Civil and Environmental Engineering and Engineering Systems
Thesis Committee Member

Certified by .....
Irving Wladawsky-Berger, Visiting Lecturer of Engineering Systems
Thesis Committee Member

Certified by .....
Wei Lee Woon, Assistant Professor, Masdar Institute of Science and Technology
Thesis Committee Member

Accepted by .....
Nancy Leveson, Professor of Aeronautics and Astronautics and Engineering Systems
Chair, ESD Education Committee

Table of Contents

Figures

Tables

LEVERAGING SOCIAL INFORMATION SYSTEMS: USING BLOGS TO INFORM TECHNOLOGY STRATEGY DECISIONS

Satwiksai Seshasai

Submitted to the Engineering Systems Division in June 2011 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Engineering Systems

Abstract

In the knowledge economy there is a need to develop new methods for processing Internet-based information to achieve growth. This is particularly applicable in complex systems where fact and perception of reality influence each other, and social information systems bridge the two. Social information systems are systems which allow participants to produce, share, integrate and comment on content. In making decisions within complex systems, there is often not one correct decision, and statistically oriented tools provide the ability to assess alternatives.

This study provides a set of algorithms and associated strategies for doing this, focused on blogs as a representative social information system. The algorithms automate the searching of blogs in a specific domain, and collection of time-series statistical data over a multi-year period around the co-occurrence of various terms within the domain. The process developed in this study uses the system representation phase of the Complex Large-Scale Interconnected Open Socio-technical System (CLIOS) process to interpret the statistical data. Subject matter experts comment on which specific statistical insights shall be used to adjust the system representation and thus lead to new technology strategy decisions being made.

Three case studies were conducted with in-depth blog data analysis and expert interviews – the domains were cloud computing in industry, broadband expansion in Kenya, and renewable energy study in Abu Dhabi. Each case study produced a quantitative assessment of the blog analysis, and also a set of technology strategy areas which would benefit from this research.

The intellectual contribution of this research includes enhancements to the theory of social information system around statistical analysis methods and extensions to the CLIOS process, enhancements to the study of blogs and the ability to study broader information sources, an implementation pattern for studying other domains, and enhanced technology forecasting. This research set the stage for five areas of broader impact: automatically reviewing information without reading it, gaining early awareness of insights, connecting experts and amateurs in a field, understanding public perception in a field, and building research tools for broader research.

Thesis Supervisor: Stuart Madnick
Title: John Norris Maguire Professor of Information Technology, MIT Sloan School of Management and Professor of Engineering Systems, MIT School of Engineering

Acknowledgements

This research would not have been possible without the support and guidance of a number of individuals and institutions. First and foremost, my doctoral committee provided me with wisdom at every stage of this process, from helping to structure the research agenda to helping with logistics of completing the case studies and writing the final dissertation. My committee chair, Prof. Stuart Madnick, has a vision for the role of information technology in society which helped motivate me to pursue this research and work in this field professionally. He helped me balance achieving the vision with pragmatically completing the research. Prof. Joe Sussman introduced me to the CLIOS process and also provided a level of detailed review of my work which significantly improved the quality. Prof. Irving Wladawsky-Berger helped keep my focus on the vision and impact of this work and his perspective coming from industry was invaluable. He also served as a bridge with my sponsor and employer, IBM, and provided me with significant professional endorsement inside of IBM. Prof. Wei Lee Woon was the main technical supervisor of my work and helped build the algorithms and tests used in the research. He also graciously hosted my visit to Abu Dhabi and set up a number of interviews for the energy case study. I was very fortunate to have such a well-balanced and dedicated committee and could not have completed this work without them. There are many other faculty members at MIT who have helped me over the 14 years and 4 degree programs I have completed, and the community support of professors and student colleagues is critical in completing doctoral work. I would like to acknowledge one additional faculty member, Prof. Amar Gupta, who mentored and supervised me through my undergraduate and two master's degrees, and set the stage for my doctoral admission and work in ESD. Even since moving to the University of Arizona, he has kept in touch with my work and progress, and having a mentor like him has helped keep me focused on my professional and personal goals.

IBM has been my sponsor and employer for the past eight and a half years and I am extremely fortunate to have had the support of many managers, executives and colleagues at the company. My managers over the years have not only supported the time required to pursue the masters and doctoral program, but also never hesitated to keep providing me with professional opportunities to grow inside the company. The executives at IBM have maintained support for me through very tough business circumstances and I am well aware that there are few companies who still offer this level of support. Finally, my colleagues at IBM have stepped in for me when work arose and I was at MIT, and have also been supportive in constructing the cloud computing case study and introducing me to the HorizonWatch program which helped motivate this research. In addition to IBM, I would like to acknowledge the Masdar Institute of Science and Technology, which provided some of the project funding, the professors involved in the energy case study, and also Prof. Woon's time and assistance. Finally, the case study in Kenya would not have been possible without the enormous time commitment and initial research project done by Brian Omwenga at MIT. Brian provided a structure for the case study, personally hosted my stay in Kenya and escorted me to every interview. Brian and all of the interviewees in Kenya really helped demonstrate the impact this research can have in a society such as Kenya.

While MIT and IBM have been great institutional sponsors of this work, I would not have pursued nor completed this dissertation without my family. My parents, Roopa and Seshu, have set education as a primary focus for me and my siblings for our entire lives. They taught me to keep learning in every endeavor and to pursue creative solutions when challenges arise. I will always be in awe of the sacrifices they made in immigrating to this country and they've set the bar high for the type of parent I'm striving to be. I became a parent during the course of this program, and our daughter Talia served as inspiration to me as I completed this work. I look forward to reading her dissertation one day. Finally, my wife Rebecca has provided love, support, motivation, and practical help throughout the many years I've been a student at MIT. Despite the four degrees and numerous wonderful experiences, the best gift MIT gave me was my wife.

Chapter 1 Introduction

This research is motivated by the growth of information being published to the Internet. In the industrial economy, the use of machines and novel manufacturing processes led to growth. We are currently participating in the knowledge economy, where there is a need to develop new methods for processing Internet-based information to achieve similar patterns of growth. This is particularly applicable in complex systems where both fact and perception of reality influence each other, and Internet-based information provides a bridge between the two. This study provides a set of algorithms and associated strategies for doing this, focused on blogs as a representative source of Internet-based information. Blogs are an example of a social information system. Social information systems are defined as systems which allow participants to produce content, share it with others, and allow readers of the content to provide their own commentary or integrate others' content into their own content. A significant volume of information is produced daily because of the ease of publication. We assume that no individual reader can read all of this information themselves, and that much of this information is not useful to the typical reader. Thus, new methods are required for processing this information before it can be used for a productive purpose. In making decisions within complex systems, there is often not one correct decision, and statistically oriented tools provide the ability to assess multiple possible alternatives. The core research goal of this study is to produce a repeatable methodology for using social information systems to inform technology strategy decisions.

1.1 Illustrative Example

As a simple example for this issue, consider the recent expansion of mobile platforms such as iPhone and Android. Over the last few years, mobile software application vendors have made decisions on which platforms to support. They have high level decisions such as which platforms to build their applications on. They also have lower level decisions such as which features of the platforms to leverage, what pricing plans to pursue, which geographies to pursue, and whether to focus on audio/video intensive applications or use text-based communication. These are decisions related to technology strategy. The technology strategy of a business unit is a selection of one or more technical alternatives to meet the specified business goal of the unit.

Many information sources can be consulted to help make these decisions, and blogs are one representative source. Blogs reveal the relative interest level in platforms such as the iPhone and Android, but also discuss the relationship between terms such as 3G and smartphone, or the growing interest in having integrated video cameras inside phones. Blogs may reveal that the cost of a data plan is becoming a driver for shifting to more SMS based communication and blogs may also reveal that the speed of building 3G networks in certain cities is directly related to the success of various providers in those cities. These are examples of more detailed system level characteristics which become evident to one who regularly reads blogs in a particular domain. These system level characteristics provide information to make the type of technology strategy decisions mentioned in the prior paragraph.

1.2 Methodology

The methodology developed in this research consists of a set of algorithms which automate the initial steps of reading and analyzing blogs in a particular domain, and a suggested process for interpreting the insights found during the analysis. It is not possible for a human to wake up every morning and read every blog post, news article, journal article, patent and law produced in their domain of interest. It is also not possible for any computer to process the various statistical findings and reach conclusions on the decisions described in the iPhone/Android example. We need a combination of both. We need a computer to analyze the vast amounts of available information, and provide it to a human in some digestible format so that the human's view of the domain can be adjusted to include new insights from the analysis.

The algorithms were implemented in a software tool which analyzes data from the past and uses the analysis as input to predictions about the future. This study assumes that changing levels of discussion of different topics can assist in predicting where the relative interest levels of different topics are going to head. We're seeking to use interest levels demonstrated by blogs to lead to new combinations of existing technologies, new markets in which to sell existing technology, and new ways to articulate the value of existing technology. In the future, this tool could be used to recommend investment in new technologies, but that is beyond the focus of this study.

This study has produced both a set of algorithms and a new process. The algorithms automate the searching of blogs in a specific domain, and collection of time-series statistical data over a multi-year period around the co-occurrence of various terms within the domain. The process developed in this study uses system representation as a means to interpret the statistical data in an understandable manner, and allows subject matter experts to comment on which specific statistical insights shall be used to adjust the system representation and thus lead to new technology strategy decisions being made. A system representation is an exercise conducted to produce a description of a system which provides decision makers with a tool to achieve a particular set of goals. The representation can be in a variety of formats such as paragraph descriptions or a set of diagrams which depict the key components of a system and the interactions among these components. In this study, we use a specific process developed at MIT for Complex Large-Scale Interconnected Open Socio-Technical (CLIOS) systems. The CLIOS process is described in detail later, but in summary it provides a set of steps that guide experts in generating and choosing amongst a set of strategic alternatives for a particular system. The first steps in the CLIOS process are to build a system representation, using a particular vocabulary to describe the parts of the system. Specifically, a CLIOS system representation starts with the identification of relevant subsystems, and relevant elements of these subsystems known as components. Components can be further classified into common drivers, policy levers and performance measures, and a set of institutional actors is also identified as part of an institutional sphere which interacts with all subsystems. Each of these types of components has a particular meaning discussed in a later section, but the key principle is that the system is decomposed into various types. The algorithms developed in this research seek to produce data which allows the system representation to be adjusted based on insights observed in the data. Each different type

of component in the system representation requires an algorithm for proposing candidate changes to that type, by searching in the blogs in a particular manner. For example, the adjusted representation (or, "re-representation") could identify new components as common drivers, add or remove links between components, or perform other actions similar to those done in the initial representation.

Interviews conducted in this study have revealed that organizations such as IBM have employees who regularly scan blogs in an ad hoc fashion. Although there is much research on using the Internet to tap into the "wisdom of the crowds", the state of the art is focused primarily on the collection of quantitative data from the Internet or on documenting very specific anecdotes of serendipitous interactions on the Internet. We believe this dissertation is one of the first pieces of work that uses computer-generated results to assist a human subject matter expert in producing or revising the representation of an entire system.

1.3 Research Questions and Hypotheses

To structure the research, a pair of specific research questions is proposed, followed by a set of hypotheses which build answers to the questions. The research questions being considered in this study are:

- Can one develop a method of scanning social information systems that better informs people making technology strategy decisions?
- Can improved information retrieval and representation of social information systems lead to different outcomes for technology strategy decision makers?

This research project is focused on the creation of a process rather than an exhaustive study of any specific domain. Domains were selected as case studies to further articulate the process improvement. In this study we selected cloud computing, broadband expansion in Kenya, renewable energy and intelligent transportation. The insights found in each domain are believed to be accurate but represent a subset of all possible insights for the domain. This is because they are based on topics which arose in the limited case study interviews done and they do not cover all relevant topics within each of the domains studied.

To help structure our research, a logical series of hypotheses were constructed. The tests required to validate each of these hypotheses forms the basis of the research project.

H1: Blogs posts contain information which is both of interest and useful to subject matter experts in various domains.

This hypothesis forms the basis of the project – if blogs were not of interest and useful, then this project would not be worth pursuing. We are not suggesting that all information in all blogs are useful. Instead, there are kernels of insight to be realized and people who read blogs do so in an intelligent and analytical manner. Although this may seem obvious to some, since there is not a

well established body of research using blogs as the data source, it's important to explicitly consider this hypothesis. After all, a research project could have been proposed which studied graffiti markings in a city – but this is not likely to be of interest to experts in a technical field. Studying published literature, on the other hand, is likely to be of interest. This hypothesis is simply to establish blogs as closer to published literature than graffiti. Measuring an exact level of interest is not relevant to this project, especially because the methods developed here can also be applied to other data sources such as published journals.

Testing H1 is based on stakeholder interviews. The exact steps to test H1 are:

- Define and identify domains to be studied in this research
- Define and identify subject matter experts
- Define interest (keywords which indicate interest)
- Define useful (keywords which indicate useful - should be those which indicate some action will come out of reviewing blogs)
- Conduct interviews with stakeholders
- Assess if blogs are already being read (likely in an ad hoc manner)
- Observe responses to questions around interest and usefulness

H2: The content of actual blog posts over time can be collected and summarized by using a statistical algorithm. Results with actual blog posts will be less sensitive to variation in the algorithm's parameters than results with random blog posts.

Given that blogs are worth studying (H1), this hypothesis (H2) basically established that a methodology for studying blogs is indeed possible to develop. It also focuses that methodology on collecting and summarizing via statistics, rather than other possible methods such as reading every blog post. Finally, this hypothesis ensures that the algorithm is worth using by using it with random blog content and performing a comparison.

Testing H2 focuses on specific quantitative tests with random blog data. The exact steps to test H2 are:

- Define 'summary' as a means to articulate the content of blogs in some statistical fashion (specifically, groupings of terms over time)
- Define 'robust' – basically, the resulting 'summary' must be more logical than if random inputs were given to the algorithm – various specific quantitative tests will be used
- Conduct tests, putting in simulated random data at various stages of the algorithm, and test to see the resulting summary which comes out and compare against the summary of actual blog content
-

H3: Candidate changes to a system representation can be produced in a repeatable form

Given that blogs can be studied algorithmically (H2), this hypothesis (H3) asserts that the algorithm will repeatedly produce the same results for a given domain. If we can implement the algorithms in a software tool to execute each step of the study, then we have shown that it is repeatable. The algorithms in this study produce a set of candidate changes to the system representation which may or may not be acceptable to the stakeholders in the domain. The process developed in this study describes how to interview stakeholders and identify a set of changes accepted into the system re-representation. This hypothesis is focusing on the *candidate* changes to the system representation, not on the *accepted* changes. The experts who select which candidate changes to accept may have different choices based on the same candidate changes.

Testing this hypothesis formed the bulk of the work in this project. This involved building the algorithms. The exact steps to test H3 are:

- Define 'candidate changes to a system representation' as new links, new performance measures, etc.
- Demonstrate that an algorithm can be built to produce the candidate changes
- Implement the algorithm in a software tool and demonstrate that zero human input during the running of the tool was required to produce the candidate changes (thus, this is "repeatable")
- Note that a human IS required to review and select candidate changes (and that different humans would select different candidate changes to accept into their system representation).

H4: A reasonable number of accepted changes to the system representation will create insights into the domain which were not previously considered by technology strategy decision makers. These insights will inspire a variety of different technology strategy actions.

Given that we have a means for producing candidate changes to a system representation (H3), this hypothesis (H4) considers how this could be useful. We need to show that there is insight to be found which experts are not already considering; if not, the project would not be too useful! We also need to show that this insight will motivate action, and not just be of passing interest to the experts. Determining whether these actions result in a positive or negative outcome for the experts is beyond the scope of this study.

This hypothesis is tested by a set of interviews with experts. The exact steps to test H4 are:

- From interviews, define the set of actions which a technology strategy decision maker would want to pursue (examples: investing in research, pursuing a new market, …)

- Set a threshold for “reasonable” (20%, …) – then, in interviews, technology strategy decision makers are asked to categorize the various candidate changes and one category is the set of changes which were not previously considered and which can motivate action
- Conduct a series of interviews, calculate the percentage of candidate changes which fit into this category and assess whether they are above the threshold
- Document the “actions” mentioned by interviewees and note the variation in potential actions (examples: investing in research, pursuing a new market, …)

H5: The methodology developed in this study is independent of domain choice.

To generalize the project, this hypothesis ensures that the methodology developed was not too tightly coupled to one domain. By studying multiple domains, we are confirming that this methodology can be used in future studies of a broad set of domains. We expected that the manner in which blogs can be used is different depending on the domain, but the core methodology remains valid regardless of domain.

The exact steps to test H5 are:

- Define the particular domains to be studied in this research project
- Define the different key characteristics of each domain (region, technical scope, roles, …)
- Conduct interviews as described for earlier hypothesis tests
- Document results in each domain

1.4 Methodological Advance

This dissertation represents a methodological advancement in the ability to capture large scale information on a given system and translate it into an actionable result for the system being studied. Previous efforts to capture large scale "wisdom of the crowds" have been through very coarse grained quantitative measures such as popularity or poll results. Representations have the ability to describe relationships between elements of a system, behavioral characteristics of the various elements of a system, and the overall alignment between different parts of the system. Representations can lead to action. The Internet has made it possible for anyone to share information which could provide insight into relationships, behavior or alignment, but the volume of information is such that this level of sophisticated analysis has not been previously done.

The methodological advance in this dissertation is best demonstrated by directly associating the statistics produced by the algorithm with specific adjustments which can be made to the representation for a system, as part of the “re-representation” discussed in the hypotheses. Examples of this include denoting a new relationship between two topics, designating a new measure for performance of the system, or creating a new category for inquiry in a domain. Note that we do not focus on finding new elements of a system, or on capturing the exact

dynamics between elements of a system – both of these are beyond the scope of this study. However, a key part of this study is to show that there is a lot of value to technology strategists to be derived simply by "re-representation" of the known elements of the system.

1.5 Case Study Selection

This study focuses on a set of domains as sample case studies – cloud computing, broadband expansion in Kenya, and renewable energy. These were chosen based on the interests of the interview subjects at IBM, as well as the subject matter expertise of the researcher and his advisors. Common characteristics of domains in this study include a specific technical area, a set of defined actors, and a market for the technology. The findings with respect to each domain are meant to be representative of the types of insights which could be derived with this methodological advancement, and not meant to suggest an exhaustive coverage of the domains in question. An explicit decision was made to focus on a small set of domains to allow deeper explorations of the methodological advances possible. With a small set of domains, we can also conduct deeper interviews with subject matter experts in the domain to assess the validity of the insights being suggested by the information analysis. The algorithms being used to analyze the blog search results also require iteration and refinement which is better done through focus on two domains. Future study should indeed consider other domains, to explore whether these methods work better for one type of domain than another.

By focusing on a small set of domains, we also focus on identifying the key business impact of this research. Although the goal is a methodological advance, the method should show significant business value in developing technology strategy decisions if this is to be judged a worthwhile endeavor. This was investigated by conducting a set of interviews with key stakeholders within each domain, which helped craft the initial set of components to be considered in each domain as well as flesh out the various areas of business impact.

1.6 Technical Scope of the Research Project

The technical scope of this research is described here in terms of the data being collected, the dimension of time being studied, and the notion of co-occurrence. The key unit of data being collected in this study is the number of search results for terms (known as “hit counts”) which are searched for in a web-based blog search database. Since we are trying to observe broad patterns across the entire span of published blogs, we focus on statistics related to hit counts rather than inspecting the actual substance of the blogs. For example, this research would collect the data that there were 500 blogs in March of 2009 which included the terms “cloud computing” and “bandwidth”, but would not inspect any of those blogs to determine what specific thoughts each of those blog authors had on bandwidth and cloud computing.

Collecting data over a time period is a key factor in this study. To achieve the goal of system re-representation, we observe changes over time in the statistics being collected and use these to suggest changes to the system representation as described earlier. This is done as opposed to collecting statistics on a current point in time. Blogs are uniquely positioned to provide key

information around emerging technologies because of the nature of discussion and the frequency of publication. Studying blogs over time allows early growth indicators in emerging technologies to be exposed.

The mathematical models used in this study are driven by the notion of co-occurrence of terms. Co-occurrence simply means that two or more terms appear in the same blog post. We study both the co-occurrence of domain-specific terms with each other as well as the co-occurrence of domain-specific terms with indicator terms that indicate a certain type of component such as a performance measure (a component which is used to measure the performance of the system in the eyes of the stakeholders) or a policy lever (a component which is used by institutions to impact the system). For example, "cost" is an indicator term, and if a term has a high rate of co-occurrence with "cost", it could indicate that the term impacts cost. Since cost is often used to assess the performance of a system, the term could be a performance measure. An example of a policy lever is a tax rate – by setting tax rates, a government institution can influence other elements of a system.

The following are technical assumptions for algorithm:

If terms frequently appear together in the same blog post, then they are likely to be related in some fashion.

If terms appear together more or less frequently as time goes on, then they indicate a growing or declining relationship which should be considered as part of the system re-representation.

If terms appear in co-occurrence with an indicator term, it indicates that the term may be exhibiting the behavior represented by the indicator term.

1.7 State of the Art – Using Blogs as a Research Subject

The motivation for this study comes from the goal of allowing organizations to use the vast array of information sources available on the Internet to inform them of the emerging trends related to their domain. The information sources available today could allow organizations to forecast which areas of technology require attention. The field of 'technology mining' has built a foundation for forecasting technology trends by analyzing formal sources such as patents and published literature. This foundation provides a basis for doing this with a much more dynamic and flexible system such as the blogs available on the internet. Specific technology choices in a domain can be made with full information on what is being discussed inside the organization and outside in the public. Services such as Vantage Point [Vantage Point] and the Thompson Collexis Dashboard [Thompson Reuters 2008] have demonstrated that the need and interest in cultivating knowledge found in blogs is increasing rapidly, but are focused purely on a statistical view of popularity of specific topics rather than understanding the interactions between those topics within a system.

A blog, short for "web log" is commonly defined as a web page with posts published by an author. These posts are often short and displayed in reverse chronological order. Blogging technology is becoming an increasingly popular means of publishing information to a mass audience – Nielsen's BlogPulse tracking service reports over 155 million blogs being tracked in February 2011 [Nielsen 2011]. As blogs move from personal diaries to use within research and business to share ideas, they are becoming a prime example of the growing volume of information content being produced in the domain of many engineering systems. Blogging technology has produced an information system which has social and technical aspects that can be leveraged by organizations to inform their decisions about which strategic technical alternatives to pursue. Information contained in blogs can reveal trends regarding which topics are being discussed at a growing or declining rate over time. Although this study does not pursue this direction, blogs can also inform organizations about which individuals and institutions to consult, both within and outside the organization. However, to reach this potential, we must develop an understanding of how existing methods of analyzing information can be adapted to apply to blogs, and also of how the relatively frequent rate of blog posts can help adapt system representation methods to incorporate more dynamic information content.

Academic research using blogs as a data source is very limited and focused on treating blogs as barometers for studying high level domains, whereas this research uses blogs to go deeper and analyze the system characteristics within a domain. For example, current research using blogs would study the growth of the "iPhone" versus "Android" but would not cover the interrelated aspects of the mobile phone industry, related to market dynamics, technology and other factors. Through this research we will show how to use blogs to go deeper than just the high level popularity and network analysis that is currently done on blogs.

The current state of the art focuses on these areas.

- Popularity – "Is topic A more popular than topic B" based on blog postings
- Network analysis by studying which blog topics link to which others
- Semantic analysis – basic analysis of blog content and comments to categorize them

Two key gaps in the current state of the art on blogs which will be studied in this dissertation are:

- Using blogs to go deeper than popularity of high level topics, and assess system characteristics within a domain
- Using blogs to analyze the changing dynamics of a system over time

Popularity: The Neilsen Company hosts BlogPulse.com which performs very basic searches, to the level of comparing term frequency of three terms over six months within blog posts [Glance 2004] and other examples exist of research on how blogs can predict general interest level in various topics [Kaye 2005].

Network analysis: When a blog entry is posted, the author often links entries from other blogs in the text of their entry. Readers of the blog can easily jump from one entry to other related entries simply by clicking on hyperlinks within the blog. Blog authors also often maintain a "blog roll" which is a single list of all other blogs which the author feels would be relevant to the readers of their blog. This ability to easily connect blogs to each other provides a mechanism for creating a network where a given blog serves as a node and links between blogs serve as edges. Network metrics such as connectedness have been applied to specific communities of blogs [Kumar 2005] to see how interconnected the various blog authors within a certain domain are. More recent works have introduced the analysis of comments to blog posts [Mishne 2006], since each comment includes the individual who posted the comment; this provides another way to connect blogs in building a network of blogs within a domain.

Semantic analysis: Studies on blog data either extract, index and analyze a particular corpus of data from the web, or perform queries using internet search engines across the entire space of blogs. Moving from studying the nature of blog usage to examining the content of blogs for insight, studies have looked at segmenting corporate blogs and applying known methods of text analysis such as latent semantic analysis to assess the relative interest levels of various terms [Chen 2007]. Anjewierden and Efimova combine text analytics with network theory to categorize blogs across the dimensions of people, documents, terms, links and time [Anjewierden 2006]. The Digital Organization Research Institute has gone further, and introduced the notion of Semantically-Interlinked Online Communities and developed a formal ontology for expressing the main topics within a blog [Breslin 2005]. This effectively allows categories within a domain to be constructed by observing the blogs within the domain.

Gaps: The research described above shows that blogs have not been truly leveraged to their fullest extent. If one regularly reads blogs in a particular domain, the discussion in those blogs provides insights into various characteristics of the domain. With the large volume of blogs that are published daily, it is nearly impossible for a person to keep up. That is the gap that this research is filling – *by using blogs to go deeper than high level popularity of different topics, and actually extract system level characteristics*. We've chosen the CLIOS process as one example of how to articulate these characteristics in a system representation which can be acted upon by interested parties.

The second gap this research will fill is related to time. Although the existing research covers trend analysis over time – for example, it may be possible to derive that Android based phones are gaining in popularity over iPhones – it is a much harder research problem to assess system level changes over time. If two concepts were unrelated at one point in time (such as video

cameras and smartphones) but then become linked later on, this causes changes to the system which are valuable to identify. Since new blog posts are created continuously, they provide a valuable tool for conducting these assessments of a system over time.

1.8 System Representation and the use of the CLIOS process

As mentioned earlier, system representation is an exercise conducted to produce a description of a system in a form such as a set of diagrams which depict the key components of a system and the interactions among these components. Many methods exist for producing a system representation, and in this study the CLIOS process was used. A system representation methodology provides a set of categories for the different elements of the system, and a means for describing the links between these elements.

The CLIOS Process, developed by Prof. Joseph Sussman and his team at MIT, is a process for studying Complex, Large-Scale, Interconnected, Open, Socio-Technical Systems [Sussman 2007]. A number of MIT Engineering Systems Division and Civil and Environmental Engineering theses have been based on this process and it provides a useful framework for dealing with systems such as the ones which this study will consider. The process focuses on systems with the nested complexity of a physical domain set within a sphere of institutions. Systems such as this align very well with the goal of this research, since we are looking not only at the information content itself but the social context of the information – who posted in the blog, and which institution they are representing. As described in the teaching note from Sussman et. al., the CLIOS process provides a framework for interested parties to see their viewpoint of a system in the context of the entire system [Sussman 2007]. This is another point of alignment with the research proposed here, as a key characteristic of blogs is that the cost of publishing is low so posts are often made without context.

In this study, the system representation phase of the CLIOS process is being used simply because it is the most geared towards articulating nested complexity. The CLIOS process also fits well with the selection of blogs as the data source for this research. Blog data is unstructured and a more formal system analysis methodology such as system dynamics would have been too structured to represent with information from blogs. For example, a system dynamics process identifies the specific equation governing the impact of one component on the other, wheras the CLIOS process allows the system analyst to simply identify a link between components. The researcher also has prior familiarity with the process, which facilitated the rapid creation of the algorithms required to build this research methodology .

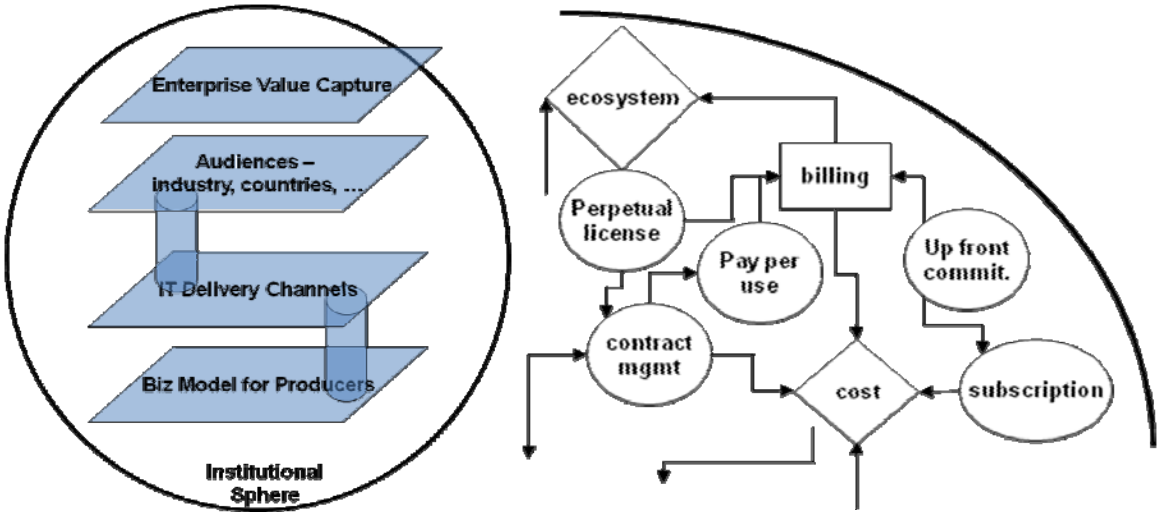

Figure 1.1 Partial CLIOS system representation diagram

Figure 1.1 shows a partial diagram of a CLIOS system representation. In this diagram, four subsystems from the Cloud Computing system are shown on the left. Each of these is described in more detail in chapter 5. For the purposes of this introduction, the key point is that subsystems represent major aspects of the system which have components that interact with each other. On the right, a small part of one of the subsystems is shown, with the interactions between components depicted via arrows. This demonstrates the concept of nested complexity because the overall system can be decomposed into subsystems but each subsystem itself has many components which interact with each other and also with components in the other subsystems.

We are not using the entire CLIOS process and would not require the surveyed stakeholders to use the entire process to select the strategic alternative their division must invest in - they may do so if they find it useful, but for the scope of this work, an actionable representation is the key requirement. An actionable representation is a representation which can lead to a particular action which would not have occurred without the representation exercise. CLIOS system representation separates technological subsystems from the institutional sphere. This is particularly useful because the core work in new methods of information extraction and analysis depend on the categorization of keywords related to various alternatives in a given domain. By separating institutions and the policy levers they use from the technical components, we can search for blogs about a particular institution and look for references to particular institutions related to particular keywords. In blog posts which have very succinct units of information, the discussion of policy levers is often distinct from a discussion of elements of the technological subsystem. Thus, the nature of our information source and our core contribution around new methods of information extraction are closely aligned with the core tenets of the system representation phase of the CLIOS process.

Another important consideration is the audience we are addressing in this research. Since we are applying these methods in the context of technology firms, we need to consider how the target audience makes decisions in their day-to-day job. At this level, individuals often spend a significant portion of focused effort on the technological subsystems (with some consideration of the policy drivers coming from institutional actors such as regulatory bodies), and then only at a periodic basis and in certain roles do they take time to look at the complexity introduced by the institutions which are affecting their particular technology domain.

The methodology developed in this research is primarily descriptive, and not normative or prescriptive as defined in the literature [Valerdi 2007]. In this research, the goal is to produce a descriptive system representation - using information extraction and analysis to describe the state of the world as it is, not as it should be. There is a significant volume of information being produced on the Internet and our goal is simply to make that volume of information accessible to decision makers by describing it in a way that it can be acted upon. It is beyond the scope of this research to speculate on the normative aspects of the representation (i.e., what topics a standard set of blog posts should cover), or a prescriptive aspect of the representation (i.e. which technology strategy decisions to pursue).

Carlos Osorio-Urzua's dissertation shows how one can extend the CLIOS system representation phase, and build a derivative model by utilizing a complementary method which provides a more refined representation [Osorio-Urzua 2007]. The study being reported here will use Osorio-Urzua's dissertation as a pattern for how to augment the CLIOS system representation phase. This is done by treating the output of the algorithm and analysis as a virtual expert which could be used just as other human stakeholders are used in the system representation phase.

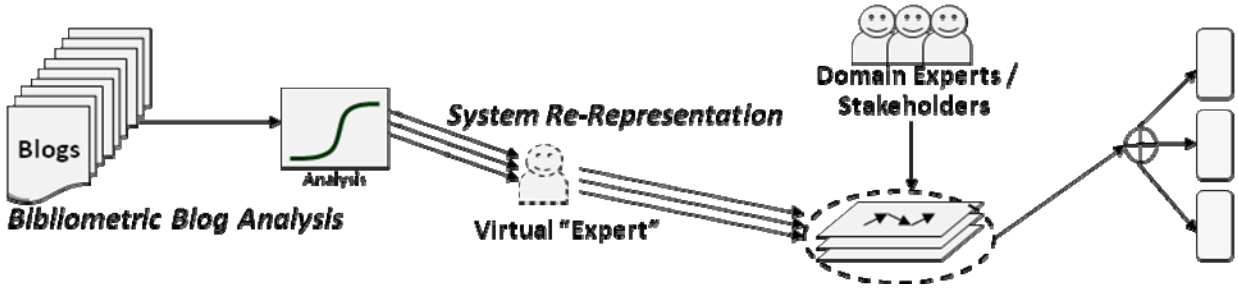

Figure 1.2 Role of the Virtual Expert in System Re-Representation

1.9

Figure 1.2 shows the role of the virtual expert in the system re-representation process being developed in this research. In the CLIOS process, human stakeholders are consulted by the system analyst to create the system representation. Just as human stakeholders are informed by numerical data, this virtual stakeholder would provide a daily update to those representing the system, and identify key trends, new topics,

individuals or institutions which may be considered in the system representation. The value of this new stakeholder is the ability to have the system representation evolve on a much more continuous basis, by using data collected by blogs to suggest new relationships between technological subsystems, or new institutions which may be considered in the institutional sphere. We assume that the reach of this stakeholder is far more global, comprehensive and more frequently up to date than human stakeholders. Visualization techniques mentioned later will help better articulate the results, just as a human stakeholder would provide a prose delivered context to quantitative data.Dissertation Structure

This dissertation will be structured as follows.

In Chapter 2, a detailed literature review is provided for each of the major phases of this study.

Chapter 3 describes the methodology which was developed in this research project is discussed. This chapter represents the major theoretical contribution of this research, with detailed coverage of the bibliometric analysis methods developed and the system re-representation process which was used.

Chapter 4 covers the assumptions made during this project. For example, certain values were assumed for parameters used in the bibliometric phase, and statistical tests were used to validate some of these assumptions.

The next three chapters 5-7 cover the three main case studies which were done in this project. Each case study chapter contains sections on the system re-representation which resulted from the bibliometric analysis, the quantitative validation done by interviewing stakeholders and asking them to categorize changes suggested in the re-representation, and the qualitative validation done by interviewing stakeholders on the overall impact of this work.

In Chapter 8 we then describe the overall results from this study, which cover patterns seen across the case studies, and suggest focused areas where this project could have particular impact.

Finally, Chapter 9 concludes the dissertation with a discussion of intellectual contributions, future study and a viewpoint of the broader context within which to view this project's results.

Chapter 2 Literature Review and Comparison to the Research

This chapter discusses the advances made to the state of the art in each of the phases of this project. Current work on bibliometrics – the collection of statistics on documents - is focused largely on comparing growth of various individual topics, and on calculating statistics related to citations of documents – it does not look at the interaction between topics over time. The goal of searching blogs over time led to an advance in blog retrieval methods made by this research to conduct period searches and also store the results in a database for future use, to circumvent a restriction by Google in the number of searches allowed. The specific analytical methods used in this research – latent semantic analysis and Google similarity distance – are among other potential methods for analyzing text based documents. The relative capabilities are discussed to justify the selection of LSA and Google similarity distance, and the application of LSA in this research is compared to how LSA has been used in other scenarios. The use of individual keywords as the primary input is a key piece of the research design and has been done by other studies.

An important piece of the methodology developed in this research is relying on human experts to select the best results among the quantitative results produced by the bibliometric analysis. The human experts in this study help bring bibliometrics into a systems context. Other studies discussed in section 2.4 have focused on a social context for the statistics rather than using the statistics to actually adjust the mechanics of a system representation. The next phase of this research project was to take the system re-representation and assess the impact through expert interviews. Other studies have tried to assess the broader impact of bibliometric analysis as well, but our combination of bibliometrics with system representation provides an advance in the breadth of impact which can be analyzed.

Table 2.1 summarizes the key areas of this dissertation and the advances in the state of the art made in each. The rest of this chapter goes into detail on the specific literature available in each area.

Area	State of the Art	Advance Made in this Project
Bibliometrics – collecting statistics on publications	Comparing growth of individual terms Network analysis of citations	Analyzing growth in the relationship between terms
Blog retrieval methods	Searching for terms in blogs Feed subscription Context extraction	Automated searches on a periodic basis combined with storing results for future use
Text analytical methods	Latent semantic analysis Latent Dirichlet allocation (LDA) Google similarity distance Taxonomy building	Application of LSA to blogs using an approximation which allows search results instead of counting the occurrences of a term in each document
Relying on human experts	Augmented experts request	Virtual expert produces insights

to interpret statistical results	specific information and receive it	which were not requested
Integration of quantitative results with qualitative process	Social context applied to statistical results	Direct mapping of the mechanics of qualitative system representation to different quantitative methods
System level impact of bibliometrics	Individual domains studied per research project	Multiple domains used in the same study to demonstrate differences in each domain as well as broader impact of this research

Table 2.1 Literature Review summary

The advances in the state of the art discussed in this chapter are on very specific independent parts of the overall research project. The major contribution of this work is in combining the overall sequence of steps – from using blogs in a domain to employing bibliometrics to system representation to interviewing experts. Together these produce a novel framework for making use of internet-based information that isn't found in the existing literature.

2.1 Bibliometrics – collecting statistics on publications

Bibliometrics is a term coined in 1969 [Pritchard 1969] used to describe the collection of statistics about published documents. A number of different methods have been developed in the literature to both collect, analyze and apply these statistics. [van Raan 1996] points out that bibliometric methods should always be used with a goal in mind, and one example of such a goal is assessing the progress of a field of research.

The statistics being collected are generally split into two categories: publication description and publication content. Publication description includes basic information related to a publication such as the title, keywords, authors, source (i.e., journal), and citations (which publications cite which other publications). [van Raan 1996] describes the use of number of journal publications in a given topic area as well as citations, and shows that even at this high level, challenges exist. Authors may have the same name, publication titles may use terms which do not clearly identify the topic, and so on. The science behind properly identifying the publication description information is a field of research unto itself. [Narin 1996] describes the use of citation statistics to indicate the influence of a certain author or certain paper on the field. This is another common area of bibliometrics, with the assumption that increasing frequency of citation indicates increasing influence.

Publication content refers to statistics that examine the terms within a publication. [Okubo 1997] describes the use of co-occurrence of pairs of words as a statistic which is used to indicate relationship between topics. An example application of co-occurrence is the construction of

maps of the different subfields of study in a broader field with relationships between subfields depicted. This dissertation builds on this area of bibliometrics by implementing a means for tracking co-occurrence of groups of words over time (not just pairs of words) and using this information to build system representation diagrams which describe relationships in various ways (instead of just based on pure co-occurrence).

This research project applies bibliometrics to the area of technology forecasting. Specifically, the focus is on building a system "re-representation" which should help a stakeholder forecast the future of their domain and use that information to pursue the right technology strategy. The formal notion of technology forecasting has existed for over forty years, described in the first volume of the *Technology Forecasting* journal by MIT professor Edward B. Roberts as the combination of predicting technical achievements and the allocation of resources towards future technological progress[Roberts 1969]. A number of methods are presented, from consulting experts to observing performance of critical metrics. One recent example is Koh and Magee's approach relying on key metrics in functional categories (such as calculations per second)[Koh 2006]. Ten years later, an article on the accuracy of various technology forecasting methods compared the goodness of fit and regression statistics of various approaches to show that simpler methods tend to be more useful [Makridakis 1979]. The 20th anniversary of Roberts' article noted the application of forecasting methods to policy makers and executives in the government, military and private sector [Hauptman 1992]. Quantitative models of bibliometric trends have been applied to business settings as well as science and technology policy settings to help assess the maturity level of a technology and when to explore a new approach [Mann 2003][De Miranda 2006]. [Courseault 2004] provides has proposed that firms should begin to look at bibliometric analysis of publication databases for competitive advantage.

The main manner in which this research project builds on existing bibliometric methods is because of the introduction of a qualitative stakeholder interview which interprets bibliometric data rather than simply using raw data to show graphs and network diagrams. The bibliometric data only produces a set of *candidate* changes to the system representation. The actual output of this phase of the research project is the system re-representation, not the raw bibliometric data. There are many factors which influence whether a stakeholder accepts a candidate change to the representation. At times, a relationship between two terms may be demonstrated by bibliometric data, and be valid, but still not included on the system re-representation because it is not relevant to the goals of the stakeholder. The same bibliometric data could very well be interpreted differently, and lead to a different system re-representation, by different stakeholders.

Another key innovation introduced by this research project is a method for analyzing the behavior of groups of terms over time. The bibliometric projects mentioned in this section all look at comparisons between terms. This research project includes a methodology for looking at whether a group of three terms, four terms, or more, grow stronger or weaker in relationship over time. This ability to expose groups of terms which all relate is very important for constructing a system representation which attempts to describe the linkages between multiple components.

2.2 Blog retrieval methods

While the existing literature on bibliometrics focuses on patents and journal articles as the typical publication sources, this research project is based on the study of blogs. Thus, a method for retrieving statistical information from blogs is required. The literature provides two main methods of blog retrieval: feed subscription and searching for terms.

Feed subscription is a technical method which relies on a common property of a blog known as a "feed." The feed is an XML representation of the latest entries in the blog, produced in an industry standard format (often RSS or ATOM). This format allows a computer program other than a web browser to "read" the entries in a blog and either store them or render them in the other program. "Feed readers" exist to allow individuals to read blogs in an application resembling an e-mail reader, rather than going to the website of each blog. A "feed fetch" is the act of consulting the feed of a blog and retrieving the blog entries currently available in the blog's feed. Many blog authors also identify a set of other blogs which they recommend, known as a "blog roll", and this provides another mechanism for programmatically collecting additional blog information. If a feed reader is fetching feeds from a blog, it can be programmed to fetch feeds from all other blogs in the blog roll. [Macdonald 2006] is an example of a blog dataset which was produced using feed subscription. In this project, 100,649 unique blogs were analyzed between December 2005 and February 2006, using 753,681 "feed fetches" (roughly once per week) to collect and store 3,215,171 blog entries in a database for analysis. [Nielsen 2006] is another example which collected 14 million blog entries from 3 million blogs (including 1.7 million blog to blog referrals). This method allows all blog content to be collected, but is limited to the number of blogs which are identified upfront and also carries a substantial cost of disk storage of the blog information which is collected. This research utilized the search method described next, and thus did not collect any blog content.

Searching for terms in blogs is an alternate approach which does not require storage of blog content but also is at the mercy of whichever blogs are stored by the search application being used. Many web sites exist which allow blogs to be searched. The key differentiator between sites is the scope of blogs which is stored. Examples of sites are:

- Blogsearchengine.com is a site which lists 150 blog search engines: Most categorize blogs and provide basic keyword search, while others aggregate feeds from specific subscribed blogs [Splashpress Media 2011]

- Technorati focuses on popularity of different blogs, based on blogs' links from other blogs [Martelli 2011]

- Blogscholar.com [Brauer 2011] and AcademicBlogs.org [Farrell 2011] are directories of academic blogs

- LexisNexis is a news search site which collects blogs which are focused on news [LexisNexis 2011]

- Google includes a separate search area for blogs – this is the search engine used for this research project [Google 2011]

The technical method to execute searches on these sites is to programmatically simulate a human user entering a search query on the website, and programmatically interpret the HTML returned by the search engine to find the number of search results. This method is described in more detail in the Methodology chapter. A more efficient method would be to utilize Application Programming Interfaces (APIs) that would allow a direct programmatic query to be made to the search engine. This capability exists for certain web search sites (such as google.com) but does not generally exist for the blog sites mentioned here.

Two advances made in this research project are to automate searches on a periodic basis and to store results for future use. Automating searches on a periodic basis refers to the fact that we searched blogs for all terms and did a full analysis of term clusters and concepts on a monthly basis, rather than just doing a one-time search or doing periodic searches for individual terms without any analysis. Storing results for future use refers to the search queries and results which were stored in a database so that future searches for the same terms and the same month did not actually go back to the original blog search site. This was done both to speed up the performance of the tool, and because Google blog search restricts the number of search queries which can be executed by a particular user on a particular day. Although both of these methods have been technically possible, the research reviewed in this area did not include any use of either of these methods.

Another key element of this research project is the actual use of blogs as a research subject. This in itself is a novel exercise and the introduction chapter of this dissertation has already covered gaps in the current literature on using blogs as a research subject. Minimal reference has been made to the use of online environments in technology forecasting [Cachia 2007] and work in this area has focused largely on the number of users who enter a website rather than inspecting content such as blogs. [Bjorneborn 2004] introduces the notion of webometrics, which refers to analyzing the links between websites. Since blog authors often include links to other blogs or other websites, this concept could very well be adapted to use with blogs, but the literature is currently focused on websites in general and not blogs specifically.

2.3 Text analysis methods

One of the critical technical aspects of this study is the text analysis method used to analyze blog content. In the prior section, we referred to two main patterns of retrieving information from blogs – either storing full blog content and studying the content, or executing searches and studying the hit counts returned by searches for specific terms. These two patterns are also

represented in the text analysis methods available to analyze blog content. Latent Semantic Analysis and additional variants such as Latent Dirichlet Allocation are methods built on analyzing stored content. When the full content is stored and available, the number of occurrences of each term in each document can be calculated. Google Similarity Distance, and additional variants such as Normalized Google Distance, are methods built on executing searches. When searches are done, the only data available is the number of documents which contain a term.

The specific steps required in each method are described in the Methodology section. The literature was surveyed on each area to help decide how to use each method in the study by reviewing other examples. The following table summarizes the two main patterns and how each was used in this project.

Pattern	**Data which can be calculated**	**Main examples**	**Use in this Project**
Store full content	Number of times each term appears in each document (term-document matrix)	Latent Semantic Analysis Latent Dirichlet Allocation	Calculating groups of terms which grow closer in relation to each other, as a group.
Search content	Number of documents which contain each term at least once.	Google Similarity Distance	Calculating the strength of relationship between pairs of terms. Calculating the relation of each term to "indicator terms" such as "cost".

Table 2.2 Text analysis methods**.**

The methods in Table 2.2 represent two dominant types of analysis, one based on the full content and one based on counting search results. Both of these methods were used in this study for producing different aspects of the CLIOS re-representation.

Google Similarity Distance refers to a measure of similarity between two terms which is calculated based on an approximation of how many websites found on Google include both terms [Cilibrasi 2007]. The specific algorithm is discussed in the Methodology chapter. This is a specific form of the generalized field of information distance [Cilibrasi 2007] which refers to the overall notion of similarity between terms. [Gligorov 2007] applies Google Distance to the notion of distance between concepts, which is calculated by defining the similarity distance between individual terms within the concept. The example used in the Gligorov study is the similarity distance between musical genres. The Google similarity distance measure has the

property of providing a quantitative weighting for relationships which allows a distinction to be drawn between relationships which should be considered or disregarded.

[Deerwester 1990] provides an overview of latent semantic analysis (LSA). A key point with this method is that it relies on having full access to the documents being studied, so that a term-document matrix can be produced. A term-document matrix is a matrix which includes every relevant term being studied, and an indication of how many times the term appears in each document being studied. This level of comprehensive analysis obviously can not be done with a search tool such as Google. Based on the term-document matrix, a set of "concepts" are produced where each concept has a weighting for each term, and this provides a view of the semantic relationship between terms. As a simple example, if four terms are often found with high frequency in the same documents, then there is an assumption that those four terms form some sort of logical concept.

In this study an approximation for the term-document matrix was used, since it was not possible to collect the full content of every blog. The approximation is discussed in the Methodology chapter. The literature contains various enhancements to latent semantic analysis which provided guidance in this study's own enhancement and approximation method of LSA. [Kumar 2006] developed a similar approximation method for reducing the computational complexity involved in a very large term-document matrix. [Farahat 2006] introduces a probabilistic element to LSA and focuses on identifying synonyms rather than related concepts. [Dasgupta 2005] introduces variable latent semantic indexing, which allows approximation based on a particular query. For example, if only a certain set of terms is of interest, the term-document matrix can be simplified to only those terms. [Chakraborti 2007] introduces adaptive sprinkling which "sprinkles" classification terms that help identify the class of usage for certain terms which have multiple meanings. For example, the term "Apple" may be sprinkled with "computer" to indicate it's being used as a description of the company rather than a description of the fruit. In addition to the extensions described above, there are also various applications of LSA which were reviewed. These include document categorization [Zukas 2003], second language development [Crossley 2008], learning [Kalz 2007], and team communications [Gorman 2003].

2.4 Relying on human experts to interpret statistical results

Although many studies use human experts to validate the results of a quantitative study, it is uncommon to use human experts to guide the intermediate steps of the study. The methodology described in this study relies on human stakeholders to develop the initial CLIOS representation, and thus select the initial terms to be used in the blog search. Human stakeholders are also used to select which candidate changes to the system representation should be included in the actual system representation.

A model for how to involve human experts in a bibliometric study is provided in [Kostoff 2000] . In the Kostoff project, the goal was to assess the frequency and relationship of technical phrases. Technical domain experts were given the set of phrases which were most frequent in the literature and asked to select which phrases were relevant for further analysis. At subsequent steps, the experts are presented with results and asked to further refine the phrases being studied. This is very similar to the approach taken in this study of using experts to select the terms in the initial system representation and then to decide which of the candidate changes to the system representation are to be included in the final re-representation. Kostoff explains that the final results are indeed subject to the biases of the technical domain expert. This is an important point because most studies in this area focus on repeatable results and on having the bibliometric results serve as an authoritative view of the domain being studied. This is one of the first projects to suggest that bibliometrics can guide the views of an expert rather than provide an complete, independent and authoritative result.

[Lancaster 1985] provides another perspective on the role of human experts in bibliometrics. The study shows that acid rain became a dominant political concern for environmentalists in the 1980s, but could've been flagged as an issue in the early 1970s if a bibliometric analysis had been done. In this case, the human expert could've been informed of potential major issues, and decided for themselves whether to pursue further action.

Another area requiring human intervention is when visualization is used to depict results. Many bibliometric studies provide statistical output and thus can be "visualized" by simple graphs. The CLIOS process utilized in this study allows for the visual depiction of the system representation to be modified based on bibliometric results. [Boyack 2002] describes a tool which allows a human expert to interact with a visualization of a domain and choose certain areas to explore further. The key point here is that even when a set of results is made available to multiple experts, the overall volume of results may be so broad that a visualization could help provide an overview and allow the expert to choose where to explore further. Different experts could each choose different areas of the results to explore in greater depth, and those choices are part of the actual impact of the study. Interactive visualization as demonstrated in the Boyack study allows for these choices to be made. In this study, a tool was used to allow experts to interact with some results, and is described in the Methodology chapter.

2.5 Use of bibliometric methods to produce system level analysis

In reviewing the literature for the various technical pieces of this project, a methodology for analyzing an overall system using bibliometrics was not found. Most studies, such as [Lancaster 1985] focus on an individual domain per study. In this study, although a single domain was studied in each case study, the multiple domains used show that the method built in this study can be generalized. The conclusion chapter of this study documents some of the differences found between the different domains in this study.

[Daim 2005] is perhaps the closest example of an integration of bibliometrics with analysis of an overall system. In the Daim study, system dynamics is used to describe the relationship between elements of the system including feedback loops. Bibliometric results were treated as an independent subsystem in the overall system dynamics model. The Daim case study was on fuel cell adoption. The main subsystems were the population, automotive, and environmental subsystems. A fourth subsystem called "maturity" was used to define the maturity of various technical alternatives and was based on bibliometric results. The output from the maturity subsystem served as input to the other subsystems. In this study, rather than treating bibliometric results as an element of the system, the bibliometric results are used to re-represent the entire system and decide which elements of the system should or shouldn't be in the representation.

This study also demonstrates that bibliometrics can be part of a broader research method involving both quantitative and qualitative methods. The results in the reviewed literature were purely quantitative in nature and focused on the characteristics of the text within the content being studied, and not on the actual meaning of the results. Mapping the bibliometric results into the aspects of the CLIOS system representation allows the interaction factors between terms in a domain to be exposed and also allows the nature of term usage to be included as a factor (for example, assessing the relationship between a term and an "indicator term" such as "cost").

The other example of bibliometric results being mapped to a system-level analysis is in the study of citation networks. [Hummon 1989 and [Small 1973] provide examples of citation networks. In a citation network, the nodes are either topics or authors, and the edges of the network are based on whether one author cites another, or whether one topic is often cited with another topic. The input to citation networks is the reference list in a journal article. If an article is about a certain topic (identified by keywords in the article's description), and it cites another article, then the topics of the two articles are deemed to be related to each other. Another approach to a citation network is to use authors or articles as the nodes. If an author or an article is cited by many others, it is deemed to be heavily influential in the field [Nerur 2005]. The statistical results of how many co-citations exist are used to depict a citation network.

Finally, [Glanzel 1996] discusses the opportunity for bibliometric results to have a broader use in scientific study and advocates for standards to help drive the field. Glanzel's assertion is that if the statistical results which come from bibliometric studies were consistent, then they could be used to drive specific decisions, in the same way that measures such as the consumer price index are calculated and used on a regular basis to drive economic decisions.

2.6 Literature Review chapter summary

This chapter provided an overview of how each phase of research extends the state of the art in relevant areas. The algorithms developed in this research have three key distinctions: they analyze groups of terms instead of individual terms, they search over time as opposed to a one-time analysis of documents, and they include both periodic automated searches as well as stored data for future searches. The combination of these three properties leads to a set of algorithms

which builds on existing state of the art in bibliometrics, text analysis, and blog retrieval. The process developed in this research for system re-representation using quantitative data and interviews with human stakeholders builds on the existing state of the art by providing a structured means for incorporating statistical data in system level analysis and then assessing the impact to stakeholders beyond simply reporting the quantitative results.

Chapter 3 Methodology

This chapter describes the algorithms and process for studying blogs which was developed in this research project. The three phases of this methodology represent the set of steps that were taken in each of the case studies.

- Phase 1: study blogs and derive insights,
- Phase 2: present the insights in the context of a system representation
- Phase 3: interview stakeholders to assess the results.

This methodology is designed in a componentized fashion so each component can be replaced by another method in future research. For example, the particular statistical methods chosen in phase 1 can be swapped with another without affecting the steps taken in phase 2 or 3. By developing this methodology and testing the assumptions made in this research project's use of the methodology, hypotheses 2, 3, and 4 are addressed. These hypotheses are repeated below:

H2 (addressed in phase 1): The content of actual blog posts over time can be collected and summarized by using a statistical algorithm. Results with actual blog posts will be less sensitive to variation in the algorithm's parameters than results with random blog posts.

H3 (addressed in phase 2): Candidate changes to a system representation can be produced in a repeatable form

H4 (addressed in phase 3): A reasonable number of accepted changes to the system representation will create insights into the domain which were not previously considered by technology strategy decision makers. These insights will inspire a variety of different technology strategy actions.

3.1 Background – CLIOS system representation

Before describing any of the research phases, it is important to understand the mechanics of the CLIOS system representation process. In the CLIOS process, the first set of steps is around creating a system representation. There are other steps which follow, but for the purpose of this research project, we are only concerned with the system representation. The first step in system representation is to define the goals and scope of the system being studied. This creates the boundaries within which the rest of the project is conducted. An example of a system definition is "broadband expansion in Kenya." Boundaries for the system include the regional boundary of Kenya, the technical boundaries of which technologies are to be studied, and temporal boundaries in terms of which timeframe is being studied. Once the system is defined, a set of subsystems, often around 5, are defined. Subsystems represent different technical and physical groupings of components which form a logical means for studying the overall system. There is

no one exact way to define subsystems for a given system and the stakeholders being consulted are trusted to define the subsystems in a way which helps them reach their goal.

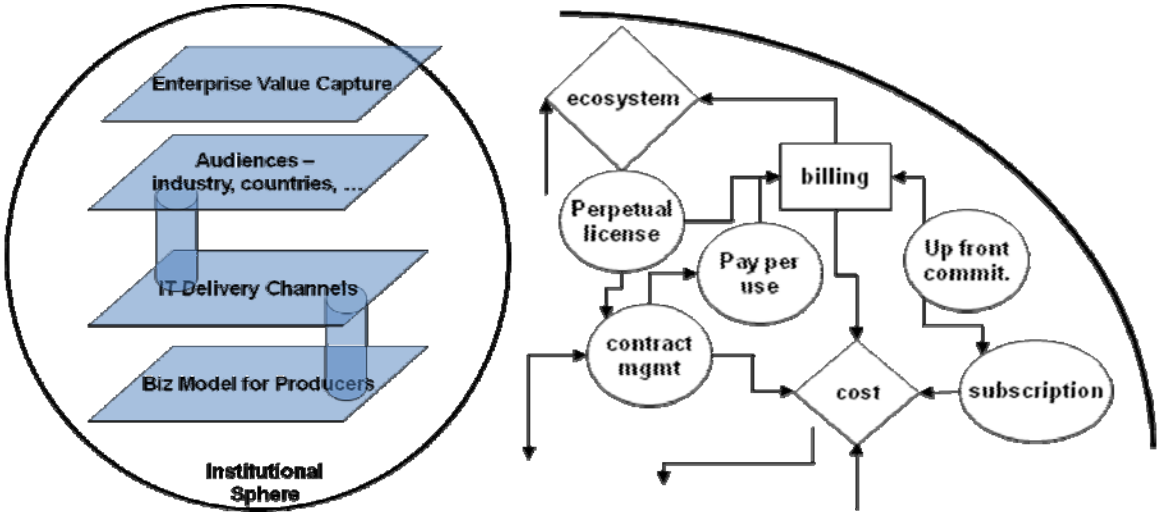

Figure 3.1 Partial CLIOS system representation diagram

Figure 3.1, the same figure as shown in Chapter 1, is an example of a graphical diagram used to show a system representation. Graphical diagrams accompanied by paragraphs and tables of text are a common way to produce the system representation. The subsystems are graphically represented as 2-dimensional planes which are surrounded by a common "institutional sphere" which contains institutions such as government bodies which affect the overall system.

The components within each subsystem are then defined, often through interviews with the various stakeholders. Most components are simply depicted as regular components (with the name of the component within a circle) but there are also three special types of components which are prescribed by the CLIOS process.

- Performance measures are components which indicate the performance of the system in the eyes of the stakeholder – for example, materials cost. They are depicted with a double line around the component.

- Policy levers are components which are used by institutions in the institutional sphere to influence the subsystem – for example, an industry regulation. They are depicted with a rectangle.

- Common drivers are components which are on multiple subsystems and thus show how one subsystem can drive another.

The graphical representation of each subsystem as a plane allows these common drivers to be depicted as cylinders which describe the way in which one subsystem can impact the others. Alongside the graphical representation, a few words of text describing the definition of each component is included.

The final step in the representation is to define the links between components. Links indicate some form of relationship between components and can be drawn as strong or weak and can also be directional in nature. The definition of a link is left purposely broad, and the accompanying text should describe the nature of each link, and what it means to be a strong or weak link.

The above elements of the system representation provide a structure by which to use the statistical results coming from the bibliometric analysis. Prior research simply produces the rates of growth of different elements and claims that one is growing faster than the other. With the CLIOS system representation framework, more fine grained results can be derived from the bibliometric analysis. This is described in phases 1 and 2.

The output of phase 1 is a set of statistical results which are intended to be used in phase 2 to produce candidate changes to the system representation. An example of a statistical result is "term A", "term B", and "term C" are growing closer in relationship in the last 15 months of the 54 month period being studied. An example of a candidate change which would come from this statistical result is "a link should be drawn between "term A" and "term C". This change would be presented back to the expert stakeholder as a candidate change, and it is up to the stakeholder to decide whether to accept the change or not.

3.2 Phase 1 – Bibliometric Blog Analysis

3.2.1 Requirements of Phase 1 – Extracting Concepts Over Time

The key goal of this phase is to identify links between terms by identifying logical concepts that exist within the overall set of terms and link some subset of the terms together. A concept is a grouping of terms which has some semantic meaning – "cats", "fish" and "dogs" may be grouped together because they are all types of domestic pets. In building the methodology, two key technical requirements motivate the approach:

1. Persistence. Concepts which are extracted must persist over some reasonable period of time. The specific terms which constitute the concept may change over time, but the underlying concept should stay persistent if it is a true concept worthy of identification.

2. Complexity: The concepts which are found need to be both interesting and relevant to the experts being interviewed. Simple pairs of terms are not sufficiently interesting as output because binary links between terms are likely to already be identified and known. Likewise, a concept which includes many terms is not acceptable because it is likely not

to represent an actual concept with semantic meaning but instead just indicates a weak relationship amongst a large group of terms.

3.2.2 Overall Process Overview

The overall goal of this phase of the research is to develop an algorithm for searching blogs in a certain domain, collect the result of these searches, and, in combination with expert inspection of the results, propose a set of candidate changes to the original system representation. The assumption is that an initial CLIOS system representation has already been created via a series of interviews with key stakeholders and analysis of other materials, and that this representation already adequately represents the domain being studied. Once this system representation is generated, the bibliometric blog analysis done in this phase will propose certain changes to the representation based on insights seen in the blogs. A candidate change could be something like "component X" and "component Y" have recently developed a strong link, or "component Z" is becoming a policy lever in the technical subsystem. The stakeholders will be presented with these changes and the hypotheses of this research project will be judged based on their reaction to the proposed changes. Figure 3.2 shows the overall process flow. The next three sections cover the three steps in the process, and subsequent sections cover the detailed implementation of the algorithm and challenges in the implementation.

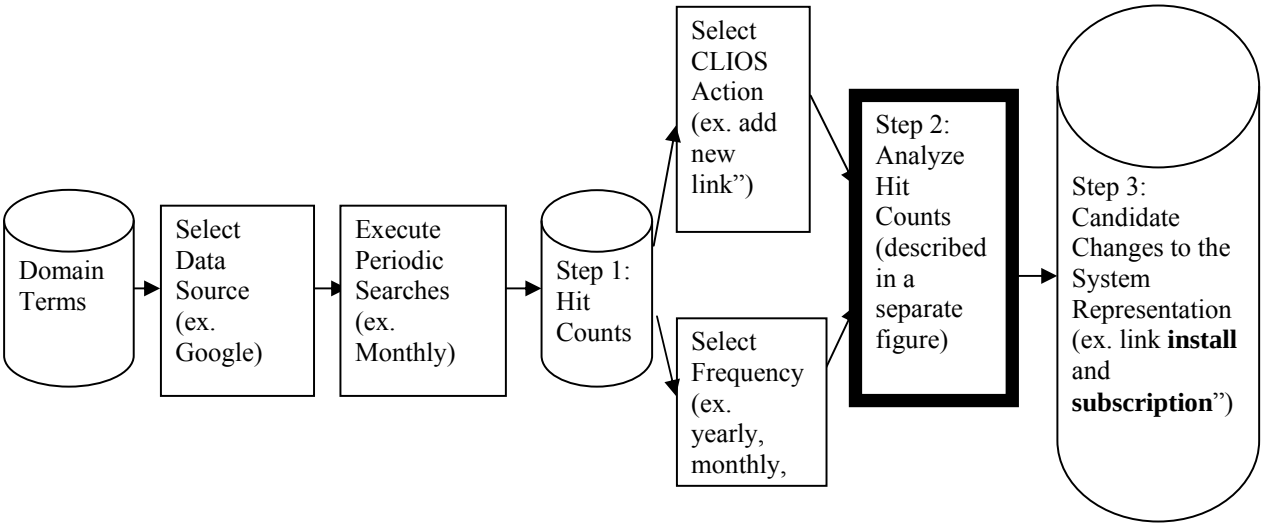

Figure 3.2 Bibliometric Blog Analysis – Process Overview. A diagram of the overall process flow, as described in the following sections.

3.2.3 Step 1 – Produce Raw Hit Counts

The first step of this process is to actually produce the raw data which will be analyzed to suggest the candidate changes to the system representation. A CLIOS system consists of

multiple subsystems. For example, the Cloud Computing system representation done in this study had four subsystems – Business Models, Enterprise Value Capture, Audiences, and IT Delivery Channels. This entire process will be repeated for each subsystem. First, a list of domain keywords is collected by taking the components of the subsystem and using them as keywords. The future study section includes a discussion of expanding this list of keywords to include related keywords through various forms of term discovery.

Given the list of keywords for a subsystem, the next step is to select a data source. For this project, the Google blog database was used. Other blog sources were considered and will be discussed in the future study section. The main consideration in selecting a data source is the scope of content – pre-selecting a set of blogs or using a more professionally focused source such as Lexis-Nexis has certain benefits, but also risks leaving out new blogs which may not have been considered before. We scope the set of blogs being considered by including the name of the overall system (ex. "cloud computing") and the subsystem ("business models") in every single search query executed, so that the only blog entries considered are those which mention both of those terms alongside the particular keywords being searched.

The next step is where most of the actual data is collected – by executing periodic searches. The goal here is to do a comprehensive search for all of the keywords over a certain span of time – for example, the past 5 years – in an effort to see which trends are occurring over recent time. To determine the right frequency to use, searches were conducted on a yearly, quarterly, monthly and weekly basis, and it was found that the monthly search frequency was the most effective for purposes of this study. This determination was made after searching for terms in the renewable energy domain and assessing the rate of growth of each term. The relative rates of growth were inspected for each frequency, and the results for the monthly searches exhibited the most reasonable fit along the growth curve. The choice of monthly frequency is effectively a design assumption of the study without any rigorous validation beyond this pilot study. The code can be configured to specify which period of frequency is used, so future study can explore further whether there is a more appropriate frequency to use.

For each month, we search for the hit count (a "hit count" is the number of results, or hits, for a particular search) of each keyword (alongside the overall system name and the subsystem name), and we search for the number of hits for every pair of keywords. For example, "install" and "subscription" were two keywords in the Business Models subsystem of the Cloud Computing domain. So, we conducted a search for the number of blog entries in the month of January 2005 which contained {"cloud computing", "business models","install"} and {"cloud computing", "business models","subscription"} and finally {"cloud computing", "business models","install", "subscription"}. Then, we did the same three searches for February 2005, and March 2005, and so on. We followed this same pattern for every single keyword, and every pair of keywords, in every subsystem. This produces a large data repository of "hit counts" that we can use for the other parts of the study.

3.2.4 Step 2 - Analyzing Hit Counts

Now that the large data repository of hit counts has been produced, the next step is to analyze these hit counts. The idea here is to apply a particular analytical method, such as Similarity Distance, Term Frequency or Latent Semantic Analysis (LSA), for a particular "re-representation action". In the Dynamic System Re-Representation section below, we discuss a distinct set of re-representation actions which can be taken against an existing system representation. These actions map directly to those specific steps one takes in building the initial system representation – identifying components, links, policy levers, etc. Each action also is mapped to a specific analytical method. The table mapping re-representation actions to analytical methods is shown below. The idea of using multiple analytical methods is to use the particular analytical method which is best suited to a particular re-representation action and also to not have the overall research project rely solely on the merits of one analytical method.

Relevant CLIOS re-representation	Analytical Method	Details
New components identified	Term frequency	Terms which increase in relative frequency over time are inspected to see if they should be new components in the system. Lowest priority among all methods.
New link between existing components	LSA	Since links may not be known, using 'concepts' from LSA (where terms can be in more than one concept) will suggest new links between components
Increased strength of link between existing components	Google similarity distance	Build clusters (with exclusive membership) among components, calculated on monthly basis, watch for new entrants into clusters
Existing components become performance measures	Google similarity distance Term freq.	Use keywords which indicate that these terms are being assessed to gauge the performance of the system Increasing frequency also inspected for perf measures
Existing components appear on other subsystems and become common drivers	Taxonomy building	Asymmetric tree of terms identified (subtopics, …). Provides a structured means for identifying whether components are appearing under other topics, over time (suggests they are becoming common drivers). "Fuzzy tree" concept of topic being in multiple nodes is also possible.
Policy levers identified with existing or new institutions	Google distance	Associate certain key terms that indicate policy levers Include institutions in all cluster creation
Re-grouping into new subsystems identified	Taxonomy building	If components are within a subtopic which doesn't align with the selected subsystems

Table 3.1. CLIOS re-representation actions mapped to Analytical Methods**.**

There are two main analytical methods which were exercised in the actual case studies – latent semantic analysis and similarity distance.

3.2.5 Step 2a - Analytical Method - Latent Semantic Analysis

Overall steps:

1. Extract concepts for each month / each subsystem – this step involves actually searching and analyzing the results for each month using LSA. In this step, there is a set of concepts produced for each month, and every concept includes a weighting for each term.

2. Parse "real concepts" at the max gap between values. The goal of "real concepts" are to select the few terms in a given concept which represent the actual underlying meaning of that concept. Since all concepts contain all terms, this step is focused on selecting and keeping only those terms which are part of the "real concept".

3. Identify concepts which grow/decay over time by indexing each concept by each pair of terms contained in that concept. If a concept has 4 terms in it, the concept is indexed 6 times – once for each possible combination of pairs of terms. The reason for doing this is that it allows pairs to be tracked over time, as this is the only method available for linking concepts over time. Thus, if in February there is a concept with 3 terms, and in March there is a concept with 4 terms, but both concepts contain a common "anchor pair", it can be said that the two concepts are the same and the third term in February and the third and fourth terms in March are considered to be associated with the anchor pair.

4. Select every concept with a pair that exists in at least 50 months (or some other threshold). Pairs which occur in 50 months or more are deemed to be worthy of presentation to the stakeholder, regardless of the size of the concepts in which the pairs appear.

5. Select every concept that exists in at least 20 months, but with less than 5 terms in each concept (or some other threshold). By constraining the number of terms in the concept to 5 or less, we can assure that each concept has some true underlying meaning. However, when concepts are constrained to 5 terms or less, the earlier constraint of 50 months has to be relaxed to a much lower threshold because the goal here is to include those concepts which could have only existed for the beginning or end of the overall time period.

6. Prepare results for presentation to stakeholder by replacing term IDs with names of terms and making any other changes required based on the software implementation of the algorithm.

The steps are discussed here in more detail. The goal of these steps is to identify new links between existing components. Our implementation of Latent Semantic Analysis (LSA) is based on an LSA simplification algorithm which relies on whether a term is in the blog post or not, rather than relying on how many times the term appears in the blog post [Ziegler 2009] . LSA takes a pre-defined set of keywords and produces a set of "concepts" where each concept contains every keyword, and a weighting for each keyword. For example, analyzing the three terms {dog, cat, bone}, might result in these three concepts: {dog 0.9, cat 0.8, bone 0.2},{dog 0.7, cat 0.3, bone 0.6},{dog 0.2, cat 0.2, bone 0.2}. Each of these concepts includes all three terms but in the first concept, dog and cat are weighted much higher than bone, and in the second concept dog and bone are weighted much higher than cat. This would suggest that there could be

a link between dogs and cats (as both are terms that describe an animal), and between dog and bone (since dogs often enjoy chewing on bones). These weightings would not indicate a strong link between cats and bones, which makes intuitive sense. In our actual study, a subsystem often has anywhere from 20 to 30 keywords and, hence,20 to 30 concepts per month. With 54 months studied in our project, each of these concepts would have to be analyzed for each month, and the strongly linked keywords detected and extracted. After the links for each month are identified, the overall set of links is analyzed to see which links emerge over time and which links disappear over time, and this produces the set of candidate links to be proposed to the system re-representation.

The standard LSA algorithm relies on a term-document matrix which is a 2-dimensional matrix where each row is a term and each column is a document and the cells of the matrix are the counts of how many times the term appears in each document [Berry 1995]. Individual blog entries are considered "documents", for the purposes of applying LSA to this research. In this case, generating that term-document matrix would require every blog entry to be downloaded on to a local machine, and every term to be counted within each blog entry. This is both impractical and may not be the appropriate approach for analyzing blogs, since blog entries are often short and having a term appear twice rather than once may not indicate that the document is two times as relevant to the particular term. Furthermore, this approach allows web-based sources such as Google blog search to be used, greatly increasing the number of source documents which are available to the algorithm. Instead of using the standard LSA algorithm, an LSA simplification algorithm was adapted from another web-search based project [Ziegler 2009]. In this algorithm, the same term-document matrix described above is used, however the values are assumed to be only 0 or 1. In other words, we only consider whether a term is in or out of the document, we do not consider how many times the term appears in the document. The standard LSA algorithm is a matrix algebra operation which takes the term-document matrix and computes a set of concept vectors as described above in the {cat, dog, bone} example [Berry 1995]. Given the m × n term-document matrix D with m terms and n documents, the Singular Value Decomposition (SVD) of the matrix is defined as:

$$D = U\sum V^T$$

where U is the m×m matrix of eigenvectors DD^T, V is the n×n matrix of eigenvectors D^TD, and $\sum$ is an m×n matrix of square roots of the eigenvalues of DD^T.

In the above SVD, the U matrix is the matrix of concepts which we desire [Berry 1995]. If we had the term document matrix D available, we could compute the covariance matrix A with each value $A_{i,j}$ representing the covariance between terms i and j. The eigenvectors of A would produce U, the concept vectors. However, based on the simplification algorithm used in this research, it is simpler to develop the covariance matrix A directly and then calculate the concept vectors in U.

The covariance matrix is derived using the following steps of standard matrix algebra [Ziegler 2009].

(1) $A_{i,j} = \text{cov}(i,j) = E[(A_i - E[A_i]) \times (A_j - E[A_j])]$

In the above equation, $E[A_i]$ is an estimate or expected value of the number of occurrences of term i in all of the documents. The outer expectation is the overall average across all documents. Thus, the above expression for cov(i,j) becomes:

(2) $\text{cov}(i,j) = \frac{1}{n}\sum_{k=1}^{n}((A_{i,k} - \frac{1}{n}\sum_{l=1}^{n}(A_{i,l})) \times (A_{j,k} - \frac{1}{n}\sum_{l=1}^{n}(A_{j,l}))$

This expands to:

(3) $\text{cov}(i,j) = \frac{1}{n}(\sum_{k=1}^{n} A_{i,k}A_{j,k} - \frac{1}{n}\sum_{l=1}^{n} A_{i,l}\sum_{k=1}^{n} A_{j,k} - \frac{1}{n}\sum_{l=1}^{n} A_{j,l}\sum_{k=1}^{n} A_{i,k} + \frac{1}{n}\sum_{l=1}^{n} A_{i,l}\sum_{l=1}^{n} A_{j,l})$

This condenses to:

(4) $\text{cov}(i,j) = \frac{1}{n}(\sum_{k=1}^{n} A_{i,k}A_{j,k} - \frac{1}{n}\sum_{k=1}^{n} A_{i,k}\sum_{k=1}^{n} A_{j,k})$

Each value in A can only be 0 or 1 because of our simplification that all values in the term-document matrix are 0 or 1 based on whether the term appears in the document at least once. Therefore, $\sum_{k=1}^{n} A_{i,k}$ is simply the number of documents which had term i in them, and $\sum_{k=1}^{n} A_{i,k}A_{j,k}$ is the number of documents which has both term i and term j. This leads to the final equation:

(5) $\text{cov}(i,j) = \frac{1}{n}(h_{i,j} - \frac{1}{n}h_i h_j)$ where h_i represents the number of hits of term i. The value of n is supposed to be the "total" number of documents, and in this research, this value was approximated by searching for all blog entries which contained the system name (for example "cloud computing").

The final equation above was easily implemented in the software algorithm by conducting a search to calculate n and subsequent searches for each term (to calculate the h_i values) and searches for every pair of terms (to calculate the $h_{i,j}$ values). This data was processed using the LSA simplification algorithm discussed above, to produce a set of concepts for every month. Each concept contained each keyword and also the weighted value of that keyword in that particular concept. So, for example, if there were 30 concepts in a subsystem, and there were 50 months studied, then a total of 1500 concepts were produced for the subsystem.

Concept 1		Concept 2		Concept 3	
Eigenvalue	-219505.9834	Eigenvalue	23233.60597	Eigenvalue	9649.401017
-------		-------		-------	
subscription	-0.92831004	**Utility**	**0.67940731**	**Hybrid**	**-0.5379518**
open+cloud	-0.234749151	**shared+infrastructure**	**0.30761498**	**platform+as+a+service**	**0.47806043**
always+on	-0.142681151	**Infrastructure+as+a+Serv**	**0.29512772**	**database+as+a+service**	**-0.3719625**
cost	-0.124815485	**utility+computing**	**0.26841094**	on-demand	0.253706077
social+networks	-0.112020883	**on-demand**	**-0.2650247**	internal+cloud	0.24590606
Software+as+a+Service	-0.108812217	Partnerships	0.176885414	shared+infrastructure	-0.22494874
Utility	-0.067726478	installation	-0.163050917	agility	-0.181034529
installation	-0.056362355	time+to+market	-0.157497156	pay+per+use	-0.1807864
time+to+market	-0.054606057	cost	-0.155887346	billing	-0.14577867
storage+as+a+service	-0.048535291	always+on	0.147325611	storage+as+a+service	0.132096598
Infrastructure+as+a+Service	-0.040663706	Hardware+as+a+Service	-0.140668221	installation	-0.128277876
Hardware+as+a+Service	-0.035089919	social+networks	-0.13776335	Hardware+as+a+Service	-0.127109841
platform+as+a+service	-0.032960492	Software+as+a+Service	-0.096999267	Ecosystem	-0.07932368
shared+infrastructure	-0.030020858	managed+services	0.093600474	cost	0.077558757
database+as+a+service	-0.027080793	storage+as+a+service	0.086883678	pilot+usage	0.070053495
internal+cloud	-0.027022937	pay+per+use	-0.074348812	Partnerships	0.050973346
on-demand	-0.024416114	open+cloud	-0.072182925	utility+computing	0.047764879
pay+per+use	-0.02225527	platform+as+a+service	0.055987222	contract+management	0.047108123
agility	-0.020138301	agility	0.049641747	open+cloud	0.043176051
Hybrid	-0.020131366	billing	-0.044908615	Infrastructure+as+a+Service	0.036876238
pilot+usage	-0.018221477	contract+management	0.038406784	Utility	0.030650893
land+and+sky	-0.018115512	pilot+usage	-0.024339322	managed+services	0.030397309
managed+services	-0.015881471	land+and+sky	-0.022304698	always+on	0.028656088
utility+computing	-0.014982681	Ecosystem	-0.011886607	land+and+sky	0.017826394
billing	-0.009742062	subscription	-0.008411586	subscription	-0.017670839
contract+management	-0.008945885	database+as+a+service	-0.007896973	Software+as+a+Service	-0.016361345
Ecosystem	-0.008809721	perpetual+license	-0.005182735	perpetual+license	-0.015161005
Partnerships	-0.008153356	up-front+commitment	0.001871724	time+to+market	0.010405785
perpetual+license	-0.003508999	internal+cloud	0.001301768	social+networks	0.006072751
up-front+commitment	-0.000133975	walled+garden	0.000484276	up-front+commitment	0.002356072
walled+garden	-9.59E-05	Hybrid	0.000480796	walled+garden	0.000439571
Cannibalization	-6.19E-05	Cannibalization	0.000284959	Cannibalization	3.84E-05

Figure 3.3 Sample LSA concepts.

Figure 3.3 depicts a sample set of concepts produced by an LSA analysis. The terms are listed alongside their relative weightings within the concept. Terms in bold represent those terms which are actually relevant to the concept.

The next step is to walk through each of these concepts, and pull out those few keywords which were most relevant to the concept – this is essentially executing the 'dog' and 'cat' example above. With approximately 30 words per concept, though, this required an algorithmic approach rather than the intuitive approach used in the 'dog' and 'cat' example. A set of options were considered and the best approach was decided to be first ordering the keywords in a concept by their values, then identifying the largest gap between values and using this as the place to draw the line between those keywords which are in or out of the concept. In Figure 3.3, concept 2, the gap between "on-demand" and "partnerships" is the maximum gap, and so this is where the distinction was made between those terms which remained in the concept and those which were removed. This is similar to how professors sometimes put student grades in rank order and decide on "A" vs "B" designations simply looking for natural breaks in the test scores. For most of the concepts, this largest gap actually came after only the first keyword – so these concepts became easy to simply disregard for our study, since they won't help us find a link between two or more keywords!

The next step is to list all concepts found in each month (now, only including those few keywords in each concept which made it above the cut of the previous step) and then by observation, looking month by month at which concepts stayed consistent over time and which concepts either were introduced or disappeared over time. This step is to meet the requirement of persistence stated earlier. After multiple approaches, the best approach seemed to be to identify those pairs which stay consistent over time (i.e., those pairs which show up for a relatively consistent set of months over a 5 year period) and then look for those other terms which join with the pair to become an n-tuple concept. Then, both the initial pair which is considered, and the occasional terms which enter into a concept with the pair, are all considered as candidate changes to propose back to the stakeholder for inclusion as new links in the CLIOS subsystem. An example of an anchor pair is described in a later paragraph in this section.

Two sets of concepts are collected independently, and the combination of these two sets meets the requirement of complexity stated earlier. One is the set of pairs which occur in some large percentage of the months (ex. 50 of 60 months) but with no limitation on the number of terms in the concept. These are pairs which are so prevalent in the blogs that they deserve to be presented to the stakeholder. The other is the set of pairs which occur in a limited set of months (ex. 20 months out of 60) but occur within a "reasonably" sized concept such as a concept with 5 terms or less. As discussed earlier, by constraining the number of terms in the concept to 5 or less, we can assure that each concept has some true underlying meaning. However, when concepts are constrained to 5 terms or less, the earlier constraint of 50 months has to be relaxed to a much lower threshold because the goal here is to include those concepts which could have only existed for the beginning or end of the overall time period. Each case study includes examples of concepts which emerged from both types of thresholds.

The complete results for each of our studies are presented in the later section on data and results. For now, we present a sample set of results to show an example of what comes out of this particular analytical method. One pair of terms that showed up consistently in the Business Models subsystem of the Cloud Computing domain was "install" and "subscription". This was a new link which was proposed to the stakeholder and made intuitive sense because with the easy installation process that cloud computing offers (the servers are managed by the vendor and almost no install is actually required), a subscription model can be offered to clients. For this pair of terms, studied over 2005 to 2009, the term "cost" was a third term often found in triples with "install" and "subscription". This made intuitive sense as well, but was already considered in the original representation. Other associated terms which showed up less frequently were:

- "time to market" (which was already linked to "install" in the original subsystem),
- "open cloud" (which was not originally considered but made intuitive sense because an open cloud allows anyone to "install" or get started more easily),
- "hardware as a service" (which was not originally considered, and generated a new business idea for the stakeholder around offering subscription based appliances), and
- "land and sky" (which made no intuitive sense to the stakeholder).

So of the sample results above, we found a handful of new links which actually could make sense on the re-representation, one particular link which was extra-special because it resulted in a new business idea, and a few links which were proposed but either were already known or made no intuitive sense to the stakeholder. This is just one example; the full analysis shown in the later section has many more candidate changes which came from this one subsystem.

These new links (produced in phase 1 of the process), when embedded into the original subsystem (phase 2 of the process), should result in new insights for the expert. The final phase of the methodology is presenting these candidate changes back to the expert (phase 3 of the process), and observing which changes were not previously considered but were accepted – if there are enough new links that fit into this category, then this tool can be considered worthwhile. Then, as we interview the expert to see what they were able to do as a result of identifying these new links, the specific hypotheses around topics such as new market approaches to existing technologies can each be assessed.

3.2.6 Step 2b - Analytical Method – Similarity Distance

This method is used when the focus is on determining the relationship between exactly two terms. These two terms could each be components in a subsystem, or they could be one component and another term such as "cost". The goal is to find the strength of the relationship between two terms by calculating the probability that if term x appears in a blog post, then term y would appear: *p(x|y)*. This probability is based on calculating the number of hits for each term, and then the number of hits for the pair of terms. The overall number of terms is included to normalize the distance measure across the entire set of possible probabilities. The below formula is the accepted calculation for normalized Google distance [Cilibrasi 2007]. In this formula, f(x) is the number of hits for term x, f(y) is the number of hits for term y, f(x,y) is the number of hits for the pair of terms, and N is an approximation of the total number of hits.

$$\frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

The CLIOS system representation process includes an indication of strength for every link between two components. This calculation allows us to assess the strength of links in the CLIOS system over time. By calculating the similarity distance on a month to month basis between two components, we can calculate the rate of growth. If the distance between two terms decreases over time, it is an indication that these two terms are growing closer in relationship. By ranking every pair of components in terms of the rate of growth in similarity distance over time, it is possible to identify those links in the system representation which could be described as stronger.

Another aspect of the CLIOS system representation is the identification of performance measures. These performance measures are used to assess the overall performance of the system towards the goals of the stakeholders for whom the representation was created. For example, "broadband speed" may be a performance measure in the Kenya broadband system, if the stakeholders see increasing the available speed as a primary goal of the system. In this study, the concept of "indicator terms" is introduced. An "indicator term" is a term which indicates that related terms may be a performance measure. To test this method, the following indicator terms were chosen: "performance", "cost", "important", "critical", "revenue", "market share",

”effectiveness”, ”speed”, ”growth”. In this study, the similarity distance between every component and each of these indicator terms was calculated on a month to month basis. This produced a set of components which showed a stronger correlation to “cost”, to “important”, and to each of the other indicator terms. These components were proposed as performance measure in the system.

The CLIOS process also includes the identification of policy levers. A policy lever is a component which is used by a member of the institutional sphere to affect the subsystem. Thus, a growing relationship between an institution such as the Federal Communications Commission and a component such as broadband speed could suggest that broadband speed were becoming a policy lever. It is possible to use this same method for identifying policy levers, but that is left for future study.

Visualizations were used to present these results back to the stakeholders. They were chosen so that the interview time with the stakeholder could be kept manageable and so that the stakeholder could quickly scan the various results and report on which candidate changes they chose to select.

The algorithm used to implement similarity distance calculation for link strength are:

1. Calculate Google similarity distance for each pair of terms, monthly

2. Calculate similarity growth/decay over time for each pair using rate of growth against linear, exponential, Gompertz curves

3. Rank pairs by the rate their similarity changes, and identify the fastest growing pairs

4. Visualize network graph of fastest growing pairs as candidate changes to the system representation

5. Present fastest growing pairs in a visual fashion to stakeholder to select

The steps used to implement similarity distance calculation for performance measures were the same, except that instead of every term pair, the pairs consisted of every term coupled with every indicator term. A line graph was produced for each indicator term, which showed those terms which were strongly correlated to that indicator term. These terms were presented back to the stakeholder as candidate performance measures

3.3 Implementation Details

3.3.1 Automated searches for blog statistics

A Python implementation was used to submit HTTP requests to automatically generated URLs of search engines – in the initial case, Google Blog Search was used. The code takes as input a search term and retrieves the number of hits for a given query. This was extended to include blogs from Google blogs and will soon be extended to include blogs from Lexis Nexis as well. One piece of the code used is shown below as a sample:

```
# Google blogs search

def gen_googleblog_search (search_term, search_year=2007, search_monthstart =
1, search_monthend=12):

return
["http://blogsearch.google.com/blogsearch?as_q="+search_term+"&num=10&hl=en&ct
z=240&c2coff=1&btnG=Search+Blogs&as_epq=&as_oq=&as_eq=&bl_pt=&bl_bt=&bl_url=&b
l_auth=&as_qdr=a&as_drrb=b&as_mind=1&as_minm="+str(search_monthstart)+"&as_min
y="+str(search_year)+"&as_maxd=31&as_maxm="+str(search_monthend)+"&as_maxy="+s
tr(search_year)+"&lr=&safe=active",lambda x:re_func(x,"of about <b>(\S+?)</b>
for")]
```

In a future step, this code may be extended to one level of web crawling, to collect the blog post title, abstract, source blog, source author, and date for the most prevalent hits. Other web wrapping technologies such as Cameleon have been developed to specialize in the extraction of content and will be investigated[1]. In the abstract of the blog entry, it may be possible to extract any other terms in the ontology, and populate a database with this information to inform future searches.

3.3.2 Latent Semantic Analysis – technical details

The following steps outline the main steps taken to execute the LSA portion of the process.

Overall set of concepts produced
Names of components in the subsystem used as 'terms', input as a list Hit counts for each 'term', for each month, collected from Google blogs using automated searches, stored in a MySQL database for easy retrieval In case a search has been previously conducted, result from MySQL is used, to avoid searching Google again Hit counts for each pair of terms, for each month, collected from Google blogs, stored in MySQL In case a search has been previously conducted, result from MySQL is used, to avoid searching Google again For each month, concepts produced with eigenvalues. The eigenvalues are the result of the LSA computations and represent weightings for each term in each concept Concepts stored as arrays of terms and eigenvalues
'Real' concepts produced for each subsystem
Terms within each concept sorted by eigenvalue For each concept, the 'gap' between each term's eigenvalue (weighting) and the prior term's eigenvalue is calculated For each concept, maximum 'gap' is calculated Maximum gap is used to delineate subset of terms which constitute the 'real' concept 'Real' concepts for each month output as a Python array
Time-series analysis of 'real' concepts
Each concept indexed by every 'term pair' which is inside the concept Then, for each term pair, there exists a set of concepts which contain that term pair over time Pull every concept with a pair that has 50 of 60 occurrences over time Pull every concept that has 20 of 60 occurrences of concepts with less than 5 terms

	Python code replaces term IDs with names of terms to output concepts as readable strings

Table 3.2. Outline of LSA implementation technical details

3.4 Phase 2 – Dynamic System Re-Representation

The concept of "dynamic system re-representation" is that a CLIOS system representation which normally takes significant stakeholder interaction to produce may be informed hourly or daily by new information sources. The goal is not to actually change the system representation on a dynamic basis, but instead to treat the output of the quantitative analysis as a qualitative stakeholder, similar to how any other qualitative stakeholder would themselves use quantitative data to provide input. Certain domains are moving at such a rapid pace that incorporating new sources of information dynamically will be very beneficial. Information sources could reveal new institutions, new individuals, new geographies, which have interest in the particular CLIOS.

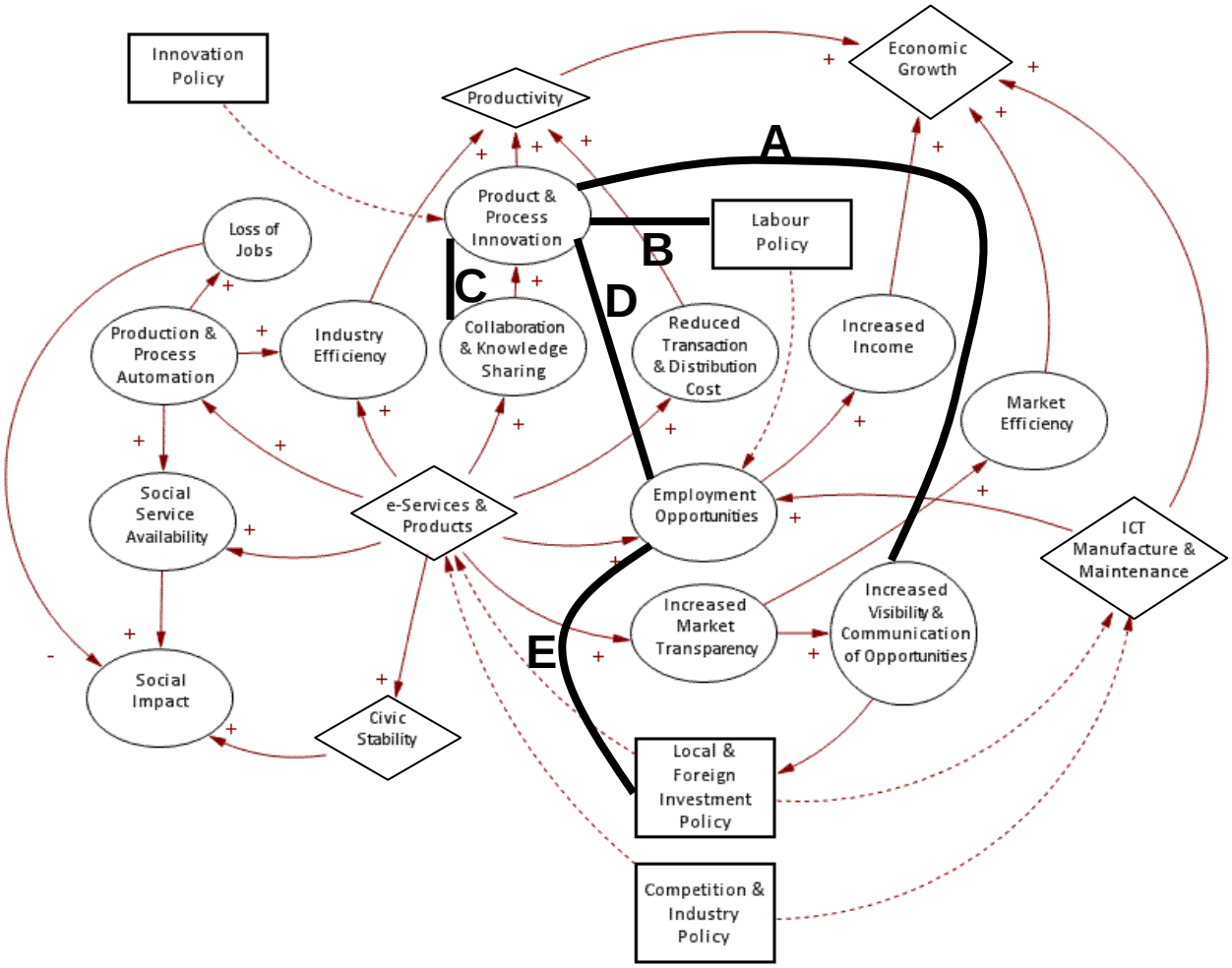

Figure 3.4. Dynamic re-representation of a CLIOS system.Figure 3.4 is another example of a CLIOS system representation. This example is from the Kenya broadband subsystem and is discussed in more detail in Chapter 6. In this figure, ovals represent regular components,

rectangles are policy levers used by institutional actors, and diamonds represent drivers common to multiple physical subsystems. Five new links are shown, which were identified through the implementation of the methodology being described in this chapter. Each of these links represents a new insight gathered from the bibliometric blog analysis and accepted by the expert interviewee as worthy of inclusion in the system re-representation. The goal of this step is to actually adjust the system representation to one which is useful to the stakeholder in making technology strategy decisions. Thus, the fact that the selected changes are based on personal opinion of the stakeholder is acceptable. When this process is used in a real situation, it is expected that the algorithm in phase 1 will be repeated on a periodic basis, so that the system representation can be re-represented to accommodate new insights being found over time.

3.5 Phase 3 - Interviewing Stakeholders to Validate the Results

The candidate changes to the system representation which were produced in phase 1 are then presented back to the stakeholder in phase 3. In this study, interviews were conducted with expert stakeholders where they were asked to select which candidate changes to the system representation they would accept into the system representation. This step is heavily dependent on the particular stakeholders being interviewed. Given the same set of candidate changes, it is expected that different stakeholders would select different candidate changes to accept into the system. This is no different than stakeholder input into the initial representation, where different stakeholders would suggest different components to be added into the representation.

In the case studies, the goal was both to test the candidate changes produced by the algorithm for usefulness as well as to understand what types of technology strategy decisions would be made based on this tool. Based on the interviews described in chapters 5-7, certain methods were found to be more useful than others in reaching insights which drove technology strategy decisions.

3.5.1 Face Validity as the Distinguishing Characteristic

The interview protocol detailed below is designed to collect expert opinion on the results which were obtained by the various analytical methods. The interview design discussed below includes two key phases – in the first phase, the experts are asked to classify each insight into one of three categories while the second phase is a more general discussion about the value of this research to the expert. The experts were asked to put insights from the research into these three categories:

1. Insights which have face validity and are already known to the expert. Face validity refers to an expert's assessment on whether an assertion presented to them has the appearance of being valid [Mosier 1947].

2. Insights which are not already known but after further exploration could have face validity.

3. Insights for which the expert is sure they do not have face validity.

By putting insights into these three categories, we were able to both assess the validity of the research method in terms of finding insights which aligned to general expert opinion as well as

finding insights which resulted in added value to the experts. In the actual interviews, the explanation of these categories took time, and did not fully sink in to any of the experts interviewed until a few insights were presented and the expert developed a familiarity with the three categories.

3.5.2 Interview Protocol

The interviews consisted of two phases. The first was a quick review of the candidate changes, asking the stakeholder to categorize each change into one of three categories as described above. The second was a more general discussion, focused on the true impact which the stakeholder saw coming from the overall set of results in this project.

Before beginning the main questions, the interviewee would describe this research project, explain the CLIOS process briefly, and go over the whole system representation and ask if there are any obvious gaps. This is done to make sure that the system representation produced by the original stakeholder was not invalid to this stakeholder. Next, each candidate change would be reviewed and categorized. This phase was often cut short because an interesting discussion around impact of the project was more critical.

In the second phase of the interview, the following set of questions was used, until the interviewee led into a particular discussion that was important to them:

- What is your intuition behind why this link showed up in the analysis?
- Does this link suggest a relationship between any existing groups within your organization that should exist but doesn't already?
- Does this link suggest a new audience for any existing products or solutions offered by your firm?
- Does this link suggest any business development activities (acquisitions, partnerships …) for your organization?
- Does this link suggest any new product ideas for your organization?
- Does the CLIOS representation help you understand the problem domain any better? If so, how?
- What is your intuition on the face validity of the new aspects to the system re-representation that were found?
- Are there any high level observations you can make about the domain given the analysis?
- Would you like to repeat this analysis on a monthly or yearly basis, or is the one time sufficient?
- Did this result in too many suggested changes to the CLIOS representation, thus making it too complicated to be useful?

After completing the interview, the results were studied and included in the case study sections that follow this chapter.

3.6 Methodology chapter summary

This chapter covered the algorithms and process which were created to effectively derive insights from blogs and present them in the context of a system representation. The algorithms for defining potential changes to the system representation such as new links between components involved a statistical collection of data from blogs and analysis using LSA and Google similarity distance. The process was defined to take the output of the algorithms, adjust the system representation, and provide the candidate changes to the stakeholders in an interview format. The interview protocol was designed to balance between a quantitative set of feedback on the validity of the results, and a more qualitative discussion of the overall impact of the results on the stakeholders' goals of making technology strategy decisions in their domain.

Chapter 4 Assumptions in the Research Design

Chapter 4 A number of assumptions were made in the implementation of this research design. It is important to note that the general pattern proposed by this research project is one which can be adjusted in many ways. At a high level, we are simply proposing that a data source be used to collect statistics about a domain over time that are interpreted in the context of a qualitative process to better understand the domain. The specific data source we've chosen (blogs), the specific analytical methods, and the specific qualitative process can each be substituted for others in a future study, without invalidating the general principles proposed by this research. However, based on the specific constraints encountered in this project, certain assumptions were made and these are described in the remainder of this chapter. Some of the assumptions were tested, and others are simply described here.

Use of Blogs as the Primary Data Source

An initial decision was made to use blogs as the data source for this research project. Examples of other possibilities include published literature, patents, legal rulings and news articles. By using blogs, we are opening a debate on the relative value of information contained in blogs. As mentioned the literature review covered earlier, blogs have yet to be used significantly in academic research and much of the reason is the uncertainty over the reliability of information in blogs versus the potential noise.

For the purposes of this project, we trust the stakeholders' own reports that blogs are already being consulted and thus there is already some initial acceptance that blogs are worth reviewing. Interviews with stakeholders in each case study confirmed that blogs were being consulted, in an ad hoc manner. At the same time, it is important to stress that the relative merit of the content of blogs is not the central question for this project. Instead, our view is that blogs provide a view of what topics are being discussed in a given domain and thus are worthy of study. Furthermore, although these is a risk that many blogs can contain invalid or useless information, the sheer volume of blogs which exist and our use of statistics is assumed to mitigate this risk.

The validation methods used in this research – interviews with experts and use of authoritative sources – do not distinguish between whether blogs are a valid data source and whether the analytical methods are valid. Future study may be used to independently verify the validity of blogs. The methods proposed in this research have been designed to be general enough to be applied even if blogs are not used as the data source.

4.2 Latent Semantic Analysis – untested assumptions

The validity of latent semantic analysis (LSA) as an appropriate analytical method is a key assumption, as the most in depth analysis in this project was done using LSA. We considered some of the pros and cons of LSA versus other possible methods such as latent Dirichlet allocation (LDA) [Blei 2003] and Kullback-Liebler (KL)divergence [Kullback 1968] and based on the simplicity of LSA as well as the ability to approximate results based on hit counts rather than requiring the full content of every document be available, we chose to conduct this part of the research using LSA. Google similarity distance was also used in another section of this project to mitigate the risk of relying on just one analytical method. Also, it is important to note that the similar analysis could be performed using any algorithm which takes a set of keywords

and produces a grouping of these terms. (examples of such methods would include any of the variety of available clustering algorithms).

4.2.1 Single Terms or Components

Within our use of LSA, many simplifying assumptions were made to conduct the research efficiently. Single terms mapped directly to components in the CLIOS representation were used, rather than utilizing synonyms or word clusters to represent each component. Google's blog search engine was trusted to conduct word stemming – searching for all forms of a word by adding "-ing", etc. A base assumption is that if terms appear together in a concept, then each term maps to a component and there is the potential linkage between those terms in the CLIOS representation. Another more extensive approach would have been to first conduct a term discovery exercise, and generate a larger list of terms for a subsystem and then use the concepts as a whole to each represent a component.

4.2.2 Maximum Gap

The mechanics of the process of going from raw results from LSA to a set of concepts to present to expert interviewees required some simplifying assumptions which were tested for robustness. LSA produces a set of concepts which include a weighting for every term, and so the assumption to use the maximum gap between weightings as the breaking point for so-called "real concepts" was tested by considering other approaches such as using a relative gap, an absolute weighting threshold, an relative weighing threshold or some hybrid combination of these approaches. Based on observation of the data in the cloud computing study, the use of maximum gap was selected, but could be further explored in future study.

4.2.3 Experts Identifying Concepts of Importance

The final assumption associated with this method is that the experts can in fact process a large number of concepts and themselves distinguish the subset of concepts which are worthy of future consideration. Another approach for future study would be to condense the results to get a more automatic representation of what is to be re-represented, and present this subset to the expert.

4.3 Latent Semantic Analysis – tested assumptions

4.3.1 Anchor Pairs

The decision to use pairs as the "anchors" for articulating changing concepts over time was tested by considering other alternatives. The LSA method returns a set of concepts for each month but the real goal of this research is to assess which concepts appear or disappear over time, and which terms appear or disappear inside those concepts. A basic assumption was made that the way to trace a concept over time is to identify an anchor set of terms that stays consistent and then track this anchor set over time to see when they appear together and which other terms appear along with them in concepts. Other options to identify and track concepts over time were not considered. The next question is how many terms are required as the "anchor" – pairs, triples, or greater "n-tuples"? Using the cloud computing and Kenya broadband data, tests were run using both pairs and triples as anchors, and the results were inspected to conclude that pairs

were the more appropriate anchor but in cases where triples resulted in concepts, those were able to be used in place of the three associated pairs.

This test was done alongside testing for other parameters as described in the next paragraph. For example, if {“cat”, “dog”, “fish”} was an anchor triple, the associated terms which arose alongside this anchor were similar to the aggregate set of associated terms for the pairs {“cat”, “dog”}, {“cat”, “fish”}, and {“dog”, “fish”}. The anchor triple {“cat”, “dog”, “fish”} may have generated associated terms of other animals such as “whale” and “dolphin”. Each individual anchor pair, such as {“dog”, “cat”} would also have generated associated terms of other animals. So, presenting the triple is more efficient than presenting each pair, because the interviewee can provide feedback in one shot as opposed to providing very similar feedback for three different but very similar pairs. Presenting the stakeholder with a triple when possible was more efficient and simpler than pairs, but pairs were required to get the full set of concepts since it’s possible that certain concepts will only result when an individual pair was used. Although in this particular example, the triple existed, in most cases, if we only looked for triples and not pairs, we would have missed many concepts which only had anchor pairs. It is our expectation that when broader n-tuples are considered, the larger anchor sets will in fact be able to replace their subsets (so, a quadruple anchor set can replace the associated triples and pairs), but this was not tested beyond pairs and triples.

4.3.2 Concepts that change over time

The next major assumption was that the best approach to presenting concepts which change over time is to split the results into two sets. The first is a set of concepts which are based on anchor pairs that occur for a large portion of the months being studied, but without regard for the size of the concepts. So, if “cat” and “dog” appear together in concepts for 50 of the 60 months, but those concepts happen to have 15 or 20 terms apiece, that is still a pair worth reporting. The second set is a set of concepts which are based on an anchor pair that occurs at some smaller threshold of months, but must occur in concepts that are limited in the number of terms. The idea here is that if “cat” and “dog only appear in 20 of the 60 months, but in those 20 months the pair appears in concepts which have 5 or less terms, then they are a pair worth reporting. Furthermore, for this second set of results, the associated terms are reported as well – so, every time “cat” and “dog” appear in a concept of less than 5 terms, the other 1 to 3 terms which appear in the concept are also reported. This method of reporting results was not tested, but the specific parameters used for each set of results were tested as described below.

The four parameters just discussed are: the “n” in the “anchor n-tuples”, the threshold for the “large” number of months, the threshold for the “smaller” number of months and the associated “term limit” for the “smaller” set. For each of these four parameters, tests were conducted over a range of possible values for those parameters, and the number of concepts produced by each combination of parameter values was inspected. Sample results are shown in Figure 4.1, and the conclusion was that a fixed set of parameters is required for each domain being studied (and must stay consistent between subsystems within a domain), but these parameter values can change from domain to domain. The selection of parameter values should be based on manual inspection of the data to find the inflection point in the number of concepts produced at varying values. The inflection point may be different for different subsystems in a domain, but we still

advocate attempting to find a consistent set of values across all subsystems in a given domain. The real mitigation for this method of parameter selection is that the expert interviews pull out the subset of resulting concepts worthy of further exploration, so if the parameters chosen produce too many concepts, the expert interviewees should be the ones who select the best subset.

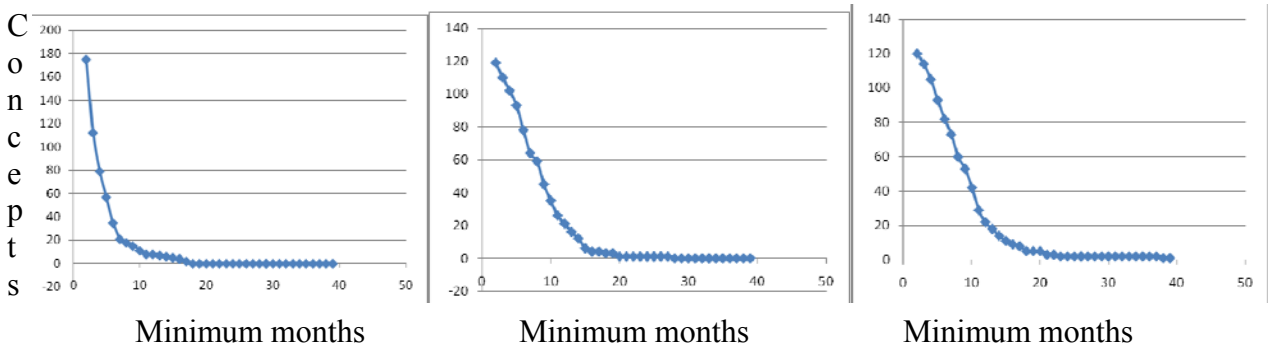

Figure 4.1 Threshold testing for "smaller" number of months.

The graphs in Figure 4.1 are for the Kenya Broadband system. Each graph represents one subsystem. Only three subsystems are shown because the curves for all subsystems exhibited similar characteristics. The x-axis is the threshold set for the minimum required number of months that a concept with 5 or less terms must appear to be counted in the output. The y-axis represents the number of concepts which were counted in the output. As the threshold for minimum required number of months goes up, the number of concepts which appear goes down. As the graphs exhibit an inflection point around 15 months, this was chosen as the threshold to use for this system. Different thresholds were chosen for the other systems, based on tests conducted for each system.

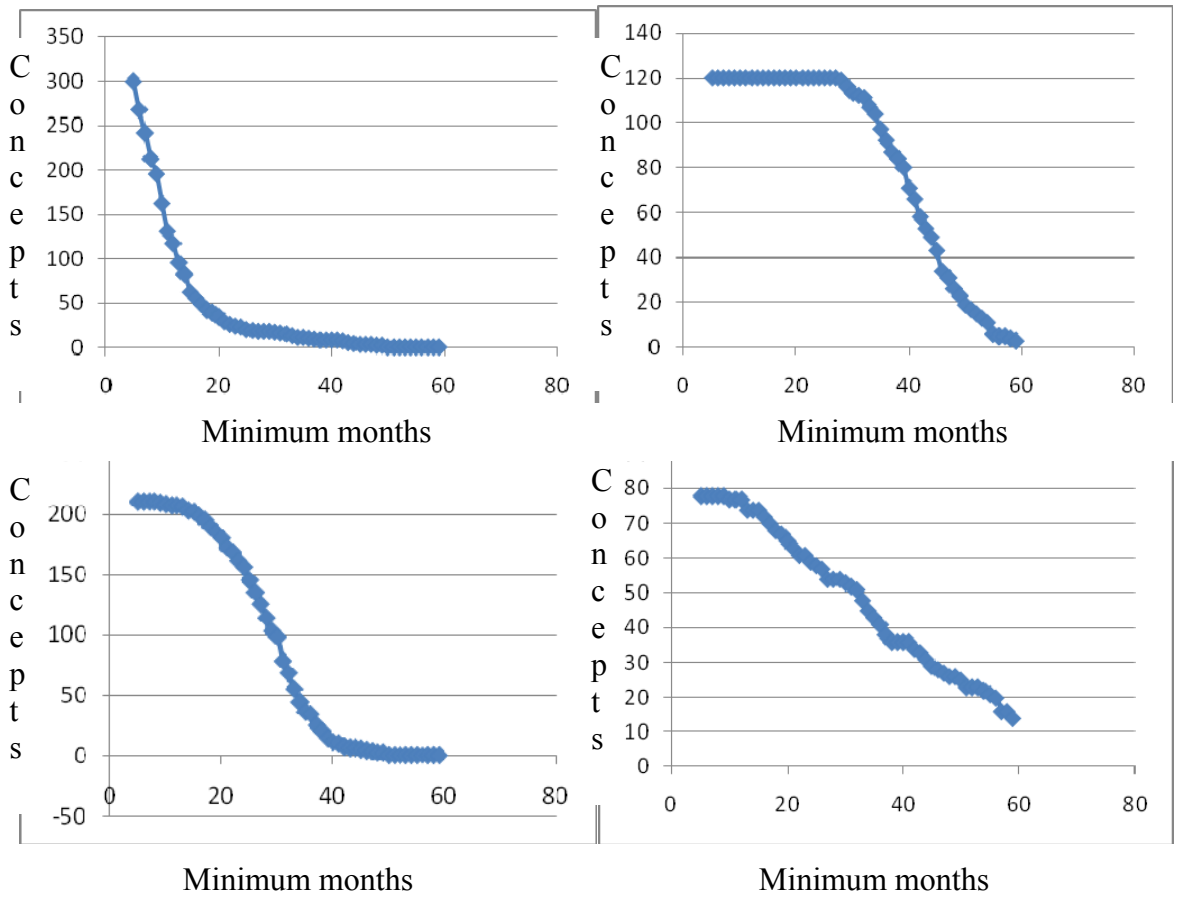

Figure 4.2 Threshold test results for "larger" number of months**.**

The Figure 4.2 graphs are for the Kenya Broadband system. Each graph represents one subsystem. In this figure, four subsystems are shown because of the differences in each of the curves between the subsystems. The x-axis is the threshold set for the minimum required number of months that a concept must appear out of a total of 60 months, *without* any limit on the number of terms in the concept. Since there is no limit on the number of terms in the concept, and there were many concepts with 10 or more terms, we expect this threshold to be quite high. If it were not high, the number of concepts in the output would have not been consumable by the stakeholder. The y-axis represents the number of concepts which were counted in the output. As the threshold for minimum required number of months goes up, the number of concepts which appear does not exhibit any regular pattern. However, since the threshold was expected to be on the high side, and since after 50 months, the number of concepts dropped to almost zero in some subsystems, 50 was chosen as the threshold.

4.3.3 Test of Randomized Concept Data

The final set of tests conducted in this area are around randomizing the concept data returned from the LSA based analysis, and then assessing whether the actual results can be distinguished from a random set of results. There were three specific tests conducted. The goal of each was to

test whether the methodology developed in this research was arbitrary or was producing statistically distinct results from randomized data.

First, the actual eigenvalues for each term in each concept was replaced with a uniform distribution of eigenvalues, and the analysis was conducted to see the number of "real concepts" produced with this random distribution of values. In a concept produced by the LSA analysis, every term being studied is given a weighting, and the terms which make up the actual concept are chosen based on the maximum gap principle discussed above. The eigenvalues represent the weighting for each term in each concept, so by randomizing them, we are testing whether the method for choosing a concept based on the maximum gap principle is valid or not. If a similar number of concepts had been produced in this test compared to the real data, then it would indicate that the maximum gap method was an arbitrary method for choosing which terms are in a concept.

Next, the table of terms and hit counts (which forms the basis for the eigenvalues to be calculated) was randomized for each month, by keeping the number of hit counts constant but varying which terms were mapped to which hit counts. This tests for consistency across the months because in the real data, a given term would be expected to slowly rise or decline in popularity, whereas in this randomized test, the mapping between terms and hit counts is changed every month. If the same number of concepts had been produced from this test, it would have been an indication that the data collected from the blogs was not actually depicting real concepts.

Finally, the mapping between terms and eigenvalues was randomized, so that in each concept the same eigenvalues were maintained but were randomly assigned to different terms. This is effectively the same principle as the previous test, but using eigenvalues instead of the raw hit counts, to further ensure the validity of the data and the method.

In each of these tests, the number of concepts produced for the subsystem was calculated with respect to a number of different random assignments, and these values were the subject of a statistical t-test to verify if the actual number of concepts falls outside of the expected number in a set of random trials.

4.4 Similarity Distance

The next set of assumptions concerns the use of Google similarity distance to calculate the changing strength of link as well as which concepts should be considered performance measures. As with LSA, the use of this method itself is a major assumption, and we acknowledge that future study may suggest that these analytical methods may be swapped in terms of which CLIOS actions they best map to.

In the specific implementation used in this project, one key assumption is the use of months as a frequency for assessing both strength of link and performance measures. While this assumption was also made in the LSA section, we believe it is more significant here because the time series aspect of the data is more prevalent. Thus, tests were conducted to change the frequency from monthly to yearly as well as quarterly. Results showed that different links emerged as the top links to consider as changing in strength. Future study is required with this method to investigate

why this was occurring, and whether the method requires further refinement to properly represent which links are increasing in strength over time.

The method of displaying results is another assumption made in this section of the research project. Two visualizations were used – the network graph and the matrix chart and these were selected as the most appropriate. Further testing could have been done to determine if different visualization techniques would have resulted in significant changes in the resulting expert opinion.

In the performance measure section of this project, the key assumption was the choice of indicator terms, especially because no synonyms were used for the indicator terms. For example, "cost" was used as an indicator term without considering other terms which also mean the same as "cost". The results did not yield an interesting set of insights for any of the indicator terms, so this entire method is worthy of reconsideration. Further study is required to test both whether other indicator terms are best or whether a fundamentally different means of analysis is required.

In assessing growth over time in each of the two uses of Google similarity distance, another assumption was that the rate of growth was normalized across the different potential links. Also, the rate of growth is obviously tied to the particular growth curve being used, and this is another assumption in this project. This project used the linear, exponential and logarithmic growth curves, and presented results to the stakeholders using the linear growth curve because this was the pattern of growth observed from the data.

4.5 Use of System Representation and Use of CLIOS

The use of system representation as a framework for expressing the statistical results was a key design decision in this research project. Given this design decision, the value of CLIOS as the system representation methodology was an assumption made without experimenting with other system representation alternatives. The core value of using CLIOS in this project is in structuring the terms and relationships between terms.

In executing the CLIOS process in each of the case studies, a simplification was made to generate the original CLIOS system representation by interviewing one stakeholder. The system representation was produced by the researcher interpreting the interview feedback of the stakeholder. There are a number of inherent assumptions in this approach which could each have been tested with further study. These are listed below:

- No bias is introduced if interviewer proposes pieces of the CLIOS system
- The explanation of CLIOS method to the expert was sufficient
- The scope and definition of domain and subsystems was appropriate
- Predefined terms are sufficient, rather than requiring the system to define new terms
- A single stakeholders is sufficient to produce a valuable experiment in this project

The Kenya broadband case study performed a partial test of the decision to use CLIOS because in this particular case study, the stakeholders had already used the CLIOS process to go through a system representation. In Chapter 6, we discuss the results of this case study and the differences found in having the CLIOS process already in place.

4.6 Assumptions chapter summary

The above sections detailed the variety of tests which were conducted to validate that the assumptions made in this research were not entirely arbitrary. Certain decisions were indeed made without full testing, and those have been documented. Other decisions were tested to some degree, but could take up an entire research project to do full testing. Overall, the interview feedback received in the case studies was sufficiently positive to indicate that the assumptions made in this project produced a methodology which is sound and acceptable.

A final thought on testing and validating the assumptions in this research project is as follows. There are 3 key technical aspects to the bibliometric analysis phase of the research – term generation, term grouping and term analysis. In our case, we have selected to use interviews to do term generation, LSA and similarity distance to do term grouping, the algorithms described earlier to do the term analysis. Each of these could be replaced by another, and the overall pattern presented here would still be a novel contribution to the field. For example, instead of interviews, terms could be generated by automated analysis of blogs in the domain. Instead of grouping via LSA, terms could be grouped by network analysis or taxonomy development. So as we explore the detailed assumptions, it's important to remember that the contribution of this work is in the high level pattern being suggested and the market-based value which is being provided to the experts being interviewed in this project.

Chapter 5 Cloud Computing Case Study

Cloud computing was selected by the IBM HorizonWatch group as one of the top ten fields which the company was interested in using in this study. Given the researcher's personal involvement in the field as well, this was selected as the primary case study for this research project. This was the first study conducted, so the process described in the Methodology chapter was developed by doing this study. Thus, certain steps were not fully automated (such as collecting the term clusters) and done with paper and pen. The evolution of the process is described in this chapter as well.

The case study follows the steps in the Methodology section, and describes how each step was conducted and the results which were found. The first section describes the initial CLIOS representation, which was developed through interviews with the HorizonWatch group. The next section describes the execution of the bibliometric analysis code, which generated a set of candidate changes to the CLIOS system representation. The next section describes the interviews conducted with five different stakeholders in the cloud computing field. Each set of interviews included an in depth review of the proposed changes to the CLIOS system representation, and then a broader discussion of how this tool could benefit the interviewee's work. The final section is a summary of the findings of this case study, both in terms of the actual results of the tool as well as the broader impact the tool has on the field of cloud computing.

5.1 Background on Cloud Computing

Cloud computing refers to the shift in computing towards more central use of IT resources through shared services to achieve scale, performance and cost. Many internet users take advantage of cloud computing services today without realizing it. When e-mail is accessed online on a service such as GMail or Yahoo Mail, this is considered using cloud computing because Google and Yahoo are using a shared set of servers to deliver e-mail services to thousands of individuals in a very low-cost manner. Before the growth of cloud computing, individuals were more likely to receive their e-mail service from a hosted provider such as their Internet Service Provider or their employer.

Other uses of cloud computing are around online services which allow software to be deployed onto a centralized infrastructure such as Amazon Web Services. Software vendors who have a particular product to make available to their customers can use Amazon Web Services to deploy their software online and take advantage of the shared investment in hardware, network, and redundancy.

Cloud computing is seen in the industry as the next evolution of computing, from mainframes to client server to internet to cloud computing. Vendors such as Google, Amazon and Microsoft each have their own perspective, and this is focused on IBM's view of the field as described by the HorizonWatch group. The field is appropriate for this study because it is a rapidly changing field, has multiple business questions to be answered, and has a significant interest in the blogosphere.

5.2 Initial CLIOS Representation

A set of interviews were conducted with two individuals on the HorizonWatch team to construct the initial CLIOS system representation. The typical CLIOS process was simplified to use a limited set of conversations, combined with presentations provided by the stakeholders, as the source material for the researcher to generate the system representation. The stakeholders were directly involved in choosing the major subsystems, and then provided significant input into the components, and described verbally some of the relationships between the components. This input was used by the researcher to generate the initial system representations and the inputs into the overall process. The simplification of using two stakeholders was appropriate because the goal of this project is to illustrate the method and not to form an exhaustive view on cloud computing. This also reduced the need of the stakeholders to fully understand the CLIOS process.

The stakeholders chose to structure the subsystems around the main business questions they sought to answer, instead of constructing a set of subsystems around the physical parts of the cloud computing system. The subsystems are shown below, depicted in the CLIOS recommended format.

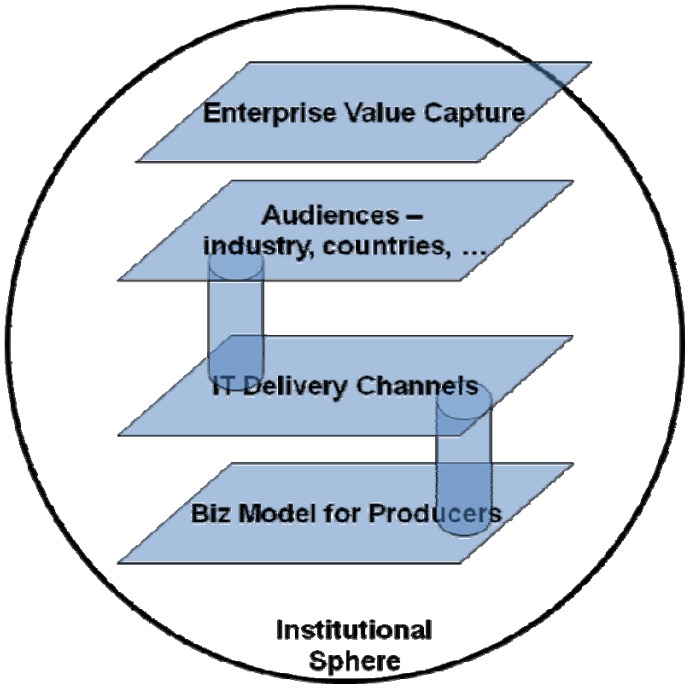

Figure 5.1 Major subsystems of the Cloud Computing system.

In Figure 5.1, each subsystem deals with a particular business question to be answered. "Enterprise Value Capture" is around how large companies can benefit from cloud computing. This subsystem involved potential ways in which an enterprise may benefit from cloud

computing, such as cost and global access to technology. The main questions to be answered are which of these benefits resonate with the public and what are the factors that can bring about these benefits. "Audiences – industry, countries" deals with specific industries or countries which may be ripe for cloud computing services. In this subsystem, countries such as Brazil and Russia were included to assess the factors impacting adoption of cloud computing services in those emerging economies. Industry specific regulations were also included to assess the relevance of cloud computing to certain industries. The main question to be answered is what factors influence each potential audience to adopt cloud computing. "IT Delivery Channels" involve the specific technical aspects of cloud computing. These are aspects such as open standards for technology, network and application security, and automation of application deployment. The main goal of this subsystem is a better view of the technical factors which are of interest to the public. The "Business Models for Producers" subsystem involves the various pricing and marketing decisions required for cloud computing vendors. This includes topics such as a subscription based model of pricing, and the various technical options for delivering IT capabilities as a service (database as a service, storage as a service, etc). This is the subsystem which would help inform stakeholders on how to go to market as an IT vendor in the cloud computing domain.

The components for all subsystems are described in table format here.

Cloud	**Enterprise Value Capture**	**Audiences - Industry, Country, …**	**IT Delivery Channels**	**Biz Models for Producers**
	long-term commitments		long-term commitments	
Common Drivers		shared infrastructure	shared infrastructure	shared infrastructure
	device access	device access	device access	
	social networks	social networks		social networks
		content anywhere	content anywhere	
Levers	open standards	communications	open standards	on-demand
	regulation	protocol	regulation	subscription
		regulation	compliance	installation
			operating system	always on
			SLA	billing
Components	scalability	developed nations	Virtualization	open cloud

labor efficiency	developing nations	automation	walled garden
business processes	mobile	web-based computing	internal cloud
business web	broadband		land and sky
dynamic allocation	data center	APIs	up-front commitment
movement of applications	bandwidth	on-demand	agility
OpenSocial	HIPPA	parallel processing	time to market
global marketplace	SOX	multi-core chips	pilot usage
cost	encryption	disk drives	pay per use
security	financial industry	distributed databases	managed services
privacy	healthcare industry	scalability	platform as a service
customization	India	flexibility	database as a service
organiational change	Brazil	security	utility computing
managed services	Russia	reliability	cost
customer relationship management	China	mashups	contract management
productivity	customer relationship management	cloud plumbing	storage as a service
			perpetual license

Table 5.1 Components in each subsystem.

The components in Table 5.1 were selected through discussion with the IBM HorizonWatch stakeholders as key areas to cover in each subsystem. They formed the terms which were input into the software tool to conduct the bibliometric analysis.

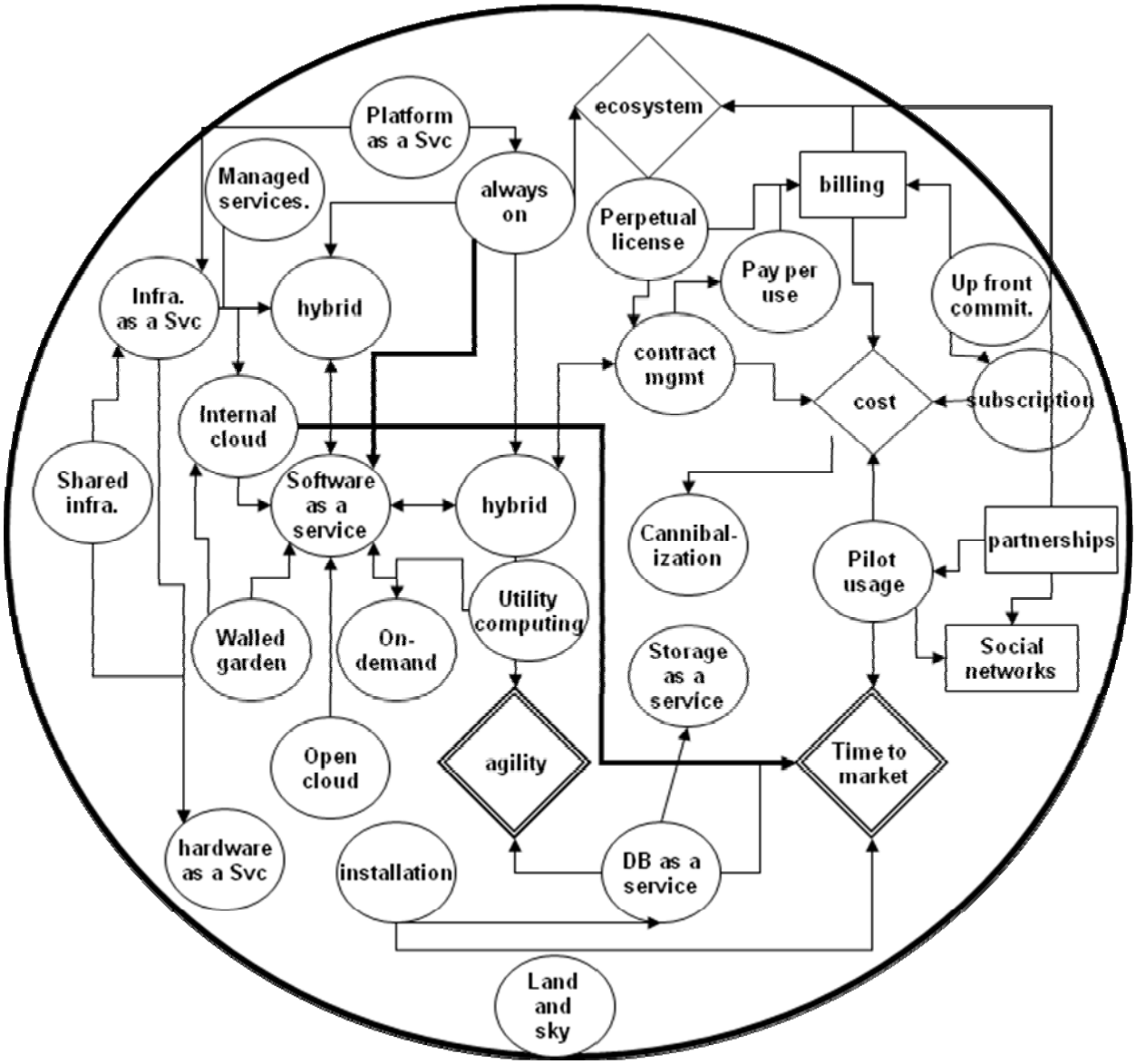

Figure 5.2 Business Models for Producers subsystem.

Figure 5.2 is a graphical depiction of one subsystem, the Business Models for Producers subsystem. The links were constructed by the researcher after discussion with the stakeholders. Once this representation was developed, this served as a basis on which to conduct the research project by running the software tool and compile a set of candidate changes to present back to the stakeholder. This representation also shows the potential impact of candidate changes – for example, if a new link is identified between two components, the stakeholder can actually assess the impact of this new link across the entire system. Examples of this phenomenon will be discussed after we review the research results.

In Figure 5.2, the Business Models for Producers subsystem is depicted using the diagrammatic tools described in CLIOS description in the introduction chapter. An example of a concept represented within the subsystem is provided here to demonstrate how this subsystem can be used to answer the underlying question of which business models are right for an IT vendor to get into cloud computing. First, the "billing" component is depicted in the top right of the subsystem as a rectangle, which indicates it is a policy lever. A policy lever is any component which can be influenced by an institution – in this case, vendors, customers or trade associations can influence how billing is conducted. There are three components with inputs into billing – perpetual license, pay per use, and upfront commitment. These three components each represent a different model for how a cloud computing service may charge their customers. A perpetual license could be a monthly or yearly fee charged as a "license" for software, a pay per use model could be used to charge based on every time a customer logs onto an online software application, and an upfront commitment could be an initial charge for a period of 3 years. The billing method is obviously different for each model, and could actually involve some elements of each. Billing has two components which it influences – cost and ecosystem. Billing obviously influences the cost of the service, as the choice in billing method would influence how much one pays for the service. Billing also has an impact on the ecosystem of partners which a vendor has – if their billing method allows them to resell cloud computing services via telecommunications providers, then their ecosystem has grown. These interaction factors are all represented in the initial system representation.

This research allows a "re-representation" of the system to occur, so when new links in the system are identified, the entire impact can be assessed. One such example is a link between "walled garden" and "upfront commitment". This link was found in the research, but is not on the initial representation. A walled garden refers to a private space for one company to use IT services, without sharing those services with other companies. A walled garden likely has a higher setup cost than a shared environment. Thus, the upfront commitment model may be the right model for cloud computing scenarios which require walled gardens. Thus, a vendor who is focused on allowing the perpetual license model for cloud computing services may examine whether their customers would accept an upfront commitment model if they were offered a walled garden. The cost may be higher, but the benefits may outweigh the costs, and the blogs which discuss this topic can provide more insight into the pros and cons seen by the market.

5.3 Candidate Changes to System Representation

After developing the initial CLIOS system representation, the next phase of the project is to develop a set of candidate changes to the system representation, using the bibliometric analysis software tool. This is the main algorithmic portion of the case study, and based on these candidate changes, a set of stakeholder interviews is then conducted to assess whether the algorithms produced a useful set of changes to the system representation.

Four types of changes to the system representation were proposed – new links, strengthened links, new common drivers, and new performance measures. The most effort was spent on identifying new links between components in the system – this proved to be the simplest type of change for stakeholders to understand and also resulted in the most in depth interview feedback. These were presented to the stakeholders in a list format which is described in the next section.

The remaining three types of changes were all presented in an interactive visual format, since they all involve some sort of relative comparison that was best navigated in a visual format.

5.3.1 New links between components

As described in the Methodology section, the goal of this section is to produce new links between the existing components in the system representation. Although a similar methodology could be used to reduce or eliminate links, that is saved for future study and not done in this case study. The method used here could potentially be used to identify new components as well, but that is left to future study. New links are identified by pairs of terms which grow closer together over time, as well as other terms which also grow closer to the "anchor pair". This method of presenting new links proved to be effective because it grouped together similar terms and allowed the interviews to focus on these groupings, rather than presenting a long list of term pairs.

In the cloud computing field, the years 2006 to 2010 were a time of significant change, both in terms of the popularity of the field as well as the change in terminology in the field. Terms such as "hosting" and "software as a service" became replaced in many situations with "cloud computing". Examples of both the change in popularity as well as the change in terminology were seen from the terms which constituted "anchor pairs" and other groupings.

A sample "anchor pair" with associated terms is shown below:

Installation-subscription, anchor pair consistently occurs through 2006-2010
- *cost*, frequently occurs with the anchor pair
- *time to market*, occasionally occurs with the anchor pair
- *open cloud*, occasionally occurs with the anchor pair
- *hardware*, occasionally occurs with the anchor pair
- *land and sky*, occasionally occurs with the anchor pair

In this sample, the terms *installation* and *subscription* (each which is a component in the initial CLIOS system representation) occur together frequently over the 2006 to 2010 time period. By"occurring" we mean the terms appeared together in one of the term clusters found in a particular month. When these two terms occur together, the term *cost* also frequently occurs. In fact, *installation* and *cost* were also an anchor pair which had many of the same other terms, and so the interviews were able to combine the discussion of this area. The other terms shown only occurred occasionally – anywhere from three to six months throughout the 2006 to 2010 time period.

A reasonable expert interpretation of this result would suggest that when people discuss the subscription based charging model which is prevalent in cloud computing, they are discussing it as a direct replacement for the need to install software on their own. *Subscription* in this context is describing a regularly fee, such as a monthly charge. *Installation* of software refers to the manual setup of software that is done when software is purchased in the traditional model. *Cost* savings are a potential benefit of this model. So these three terms are not surprises as the anchor terms, and the other terms are indeed tangentially related but not directly related. Of the other

terms, the inclusion of *hardware* (physical computers that run software) as one of the terms brought up the possibility of a new business model for offering subscription based services to hardware hosted by a service provider such as Amazon, which allow individuals to run their own software on Amazon's machines. This suggests a link should exist on the system representation between *hardware* and *subscription*. Today, that is known as "infrastructure as a service" and is a common aspect of cloud computing. However, at the time this was identified in the blogs, it was not as dominant of a model. One of the stakeholders interviewed responded that it would have aided in strategy discussions around the value of this model, and with this result, she would have consulted the specific blogs mentioning these topics and considered their viewpoints.

Open cloud refers to the aspect of cloud computing where anyone can access shared IT services such as e-mail without physical segmentation between customers' data, and *land and sky* refers to the notion that hybrid options exist where certain IT services are still maintained by the customer (*land)* and others are acquired from the cloud (*sky)*. Both of these terms were rejected by the stakeholders as not worthy of further investigation. This is an example of the type of result which came out of this study – the majority of the insights were already known to the stakeholders (thus validating the tool), but a few gems were found which generated an insight that they had not considered before but wanted to pursue further.

As this is the first case study in the project, the below table goes into detail on the raw data collected by the bibliometric analysis tool, which resulted in the set of "anchor pairs" and associated term groups being presented to the stakeholders.

July 2006	August 2006	September 2006	October 2006	November 2006
6, 12, 27	9, 14	5, 9, 15	8, 5, 9	5, 4, 20, 14
7, 9	6, 27, 12	31, 30, 20, 4	8, 19, 16, 32, 14	5, 29, 24, 4, 32, 21
9, 10, 26	6, 12, 27	30, 31	8, 16, 21, 1, 5, 10, 23	32, 15
5, 19, 29	32, 20, 4	29, 21, 19, 9, 26	5, 10, 32, 19, 21, 15	30, 31
32, 26	31, 30	27, 12	5, 1, 19	23, 11
32, 14, 21, 30, 31, 15, 10	3, 25, 17	26, 24	31, 30	20, 4, 29, 32
30, 4, 26, 32, 20	28, 18, 8	23, 32, 1, 14, 16, 18, 8	27, 12, 6	20, 24
30, 31	26, 5, 14, 1, 23, 7	21, 9	27, 12	2, 3, 11, 25, 17, 28
28, 2, 17, 21, 13	20, 4	20, 4	25, 11, 22, 17, 2, 28, 7, 3	19, 9, 15, 26, 23, 25
27, 12	16, 30, 31, 24	19, 24	24, 21	18, 26, 7, 9
21, 29	16, 24	18, 16, 10, 22, 15, 23	23, 16, 22, 28, 17, 13	17, 22, 2, 28, 9
20, 4	13, 15, 1, 8, 9, 5, 29, 22	12, 27	20, 4	17, 2
20, 4	12, 27, 26, 13	12, 27	20, 4	16, 7
2, 28	12, 27, 15, 10	11, 17, 3, 25	2, 22, 3, 25, 17, 16, 7, 23	15, 24, 23, 29, 32, 20, 14, 9, 18
18, 29, 2, 8	1, 22, 32	10, 23, 19, 8, 7, 32, 21, 14, 24, 1, 28	18, 9	12, 27
16, 24		10, 15, 25, 11, 13	18, 15	10, 32, 14, 4, 5, 21
16, 24			17, 2, 11	10, 21, 15, 20, 4
14, 31, 28, 32, 10, 19, 27, 12, 15, 1, 17, 21, 8, 23			15, 13, 11, 18	1, 8
11, 2, 19, 22, 8, 18, 28, 5, 1, 23, 15, 29, 21, 3, 14, 7			13, 28, 17, 18, 30, 31, 32	
1, 7			13, 23, 19, 9, 10, 7, 5	

Pair					
12,27	+6	+6		+6	+5,14 +29,32
20,4		+32	+31,30		+10,15,21
32,4	+30,26,20	+20	gone	gone	+20,29 +5,29,24,21
30,31		+24,16	+20,4		
1,7		group	gone	gone	[GONE]
17,24	gone	gone	gone	gone	gone
16,24		+30,31	-16 , +26,19	gone	gone
2,28	+17,21,13	gone	gone	gone	+3,11,25,17

Table 5.2 Monthly term clusters for the Cloud Computing Business Models subsystem.

In Table 5.2, each column represents a month that was studied in the cloud computing case study. Each number represents a term – this was done for execution of the software tool and the actual numbers do not indicate any ranking or other meaning. The rows in each month represent the term clusters which were found in that month. For example, the first term cluster found in

July 2006 was "6, 12, 27" which is cost (term 6), installation (term 12), and subscription (term 27). Within this cluster is the pair of terms 12 and 27, which together appear frequently throughout the entire time period. Because of this, the pair 12, 27 is designated an anchor pair (shown at the bottom of the table) and tracked throughout the months. For each month, whichever other terms appear in a cluster along with 12 and 27 are listed. In this case, term 6 appears in July 2006, August 2006, and October 2006. It does not appear with 12 and 27 for September 2006 and November 2006. This table only shows the first 5 months but in total, 48 months were studied for this case study. Since term 6 appeared frequently with terms 12 and 27, it was communicated to the stakeholders when the 12 and 27 pair was reported.

In the cloud computing case study, the above process was conducted manually as part of the effort to develop and refine the algorithm to be used. In the following two case studies, on broadband expansion in Kenya and on renewable energy, the software tool handled the process completely, and used exact thresholds for when to include terms in the final report to stakeholders. This process was described in detail in the Methodology chapter.

The total set of term clusters (and thus, suggested new links) are presented in Appendix 1. A statistical summary of the results is shown in the below table. More details on the results, plus interview feedback on the results, are presented in a later section in this chapter.

Subsystem	**Anchor Pairs with < 5 other terms (appearing in 20+ months)**	**Anchor Pairs appearing in 45+ months**
Biz Model for Producers	7	0
Audiences – Industry, Countries…	9	10
Enterprise Value Capture	11	20
IT Delivery Channels	6	7

Table 5.3. Potential new links in each subsystem.

Table 5.3 shows the number of anchor pairs found in each subsystem. The left column indicates the number of anchor pairs that appeared with clusters of less than 5 terms, and each of these anchor pairs included a set of other terms which appeared with it. An example of this is the installation-subscription term which was discussed earlier and had the cost term appear with it. So for each number in the left column, there is actually one or more links which could come out of each anchor pair. The right column describes anchor pairs that appeared in 45+ months. These are pairs which were collected without any regard for the number of other terms in the cluster. So each of these pairs would allow for just one potential new link, between the terms in the pair.

5.3.2 Strength of links

The next set of changes are around the strength of links between components of the subsystems. These results were compiled by taking the Google similarity distance between every pair of terms, for every month, and then calculating the rate of growth of the similarity distance over

time. If the strength of link between two components was found to be increasing, the stakeholder had the option of identifying the link as a strong link on the revised system representation. Those terms which decreased the most in similarity distance over the time period were depicted in an interactive network map, as shown in the below figure. The map is interactive because it was presented to the stakeholders with the Many Eyes tool discussed earlier, and stakeholders were able to drag nodes around on the screen to investigate certain areas of the map in better detail.

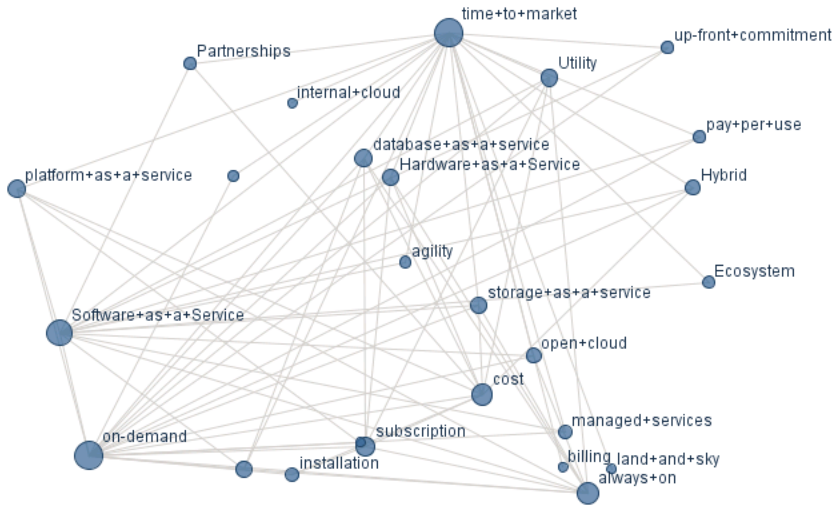

Figure 5.2 Network map of link strength in the Business Models for Producers subsystem.

In the network map shown in Figure 5.2, those pairs of terms whose similarity distance decreased the most over time are connected by an edge. The size of each component's node is based on the number of edges it is connected to. So, for example, time-to-market is a larger node because there are many components who had increasingly stronger links to it over the time period. This is an intuitive finding because while cloud computing started as primarily a cost driven activity, many have come to find that a faster time to market is another major benefit of moving towards cloud computing. This is a finding which would benefit a vendor who would've been able to realize this earlier than their competitors, especially since they could identify which specific fields are being discussed as important to time to market.

The interview feedback on this particular set of results was mixed. Although some specific findings were insightful, there is some confusion between this method and the method used to find new links, and future study will need to clarify when the right case is to use each method. Also, the specific nodes found in this map, such as the time to market node, turn out to be potential performance measures based on the interview feedback. So, it's possible that this method may be best for identifying other component types such as performance measures or common drivers. This was not pursued in greater depth in the later case studies, and the future study chapter discusses several options for exploring this further.

5.3.3 Identification of Performance Measures

Google similarity distance was also used to identify performance measures. This was done by using the same process as was done for link strength, but comparing each term to a set of indicator terms as discussed in the Methodology section. A graph was produced for each indicator term, showing how each component's distance to that term changed over time. This set of results was not found to be very useful to the stakeholders, and future study is required to further develop this method or use another method for identifying performance measures. This is likely because the indicator terms used were not broad enough to cover the actual intent of the term. For example, using "performance" as an indicator term required a blog post to include the word performance along with another term. However, there may be many ways in which a writer may choose to indicate that a given component is in fact an indicator of the system's performance. The figure below shows a sample set of results for the performance measure "price".

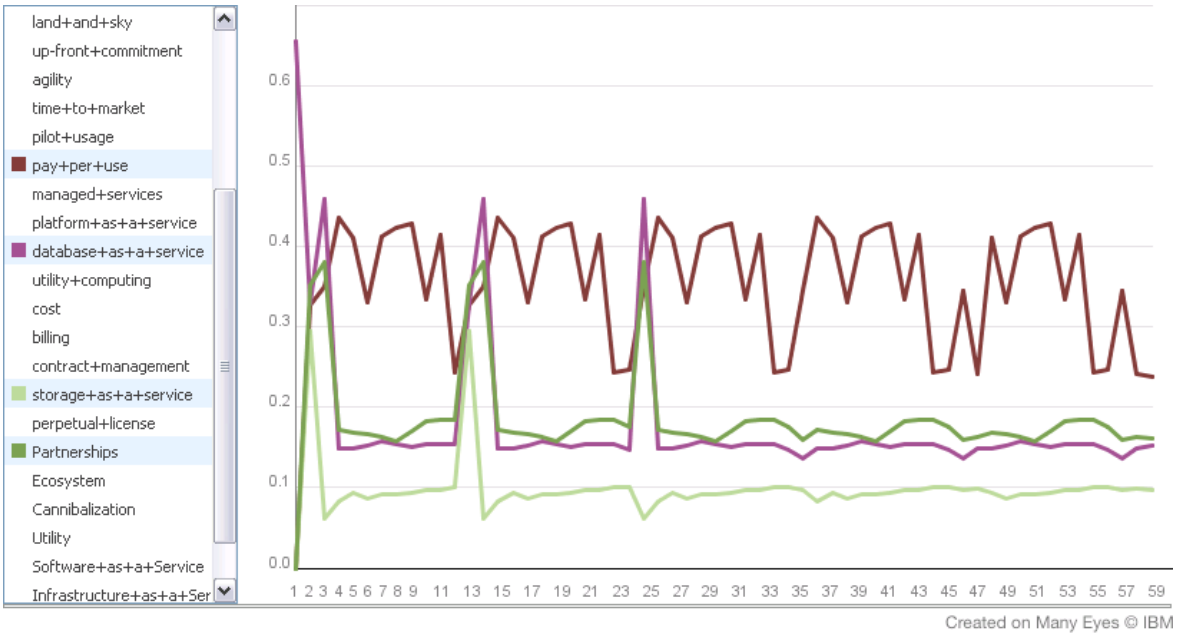

Figure 5.3 Performancemeasure analysis using the "price" indicator term**.** The X axis represents months, and the Y axis represents distance. Each line is an individual component. If the line goes down over time, or is relatively lower than the rest, it could indicate that the component could be a performance measure. However, interview feedback rejected this set of results as not being too useful.

5.4 Quantitative feedback on candidate changes

Most bibliometric studies reviewed in the literature stop here – they report on the statistical results and do not have a stakeholder review to pick which results are most useful. This case study included an in depth set of questions which produced both quantitative and qualitative validation of the results found earlier.

In this section, we will discuss the stakeholders' responses to each set of results. The interview feedback is in two sections – in this section, we cover the stakeholders' categorization of each candidate change into the three categories discussed in the Methodology section. In the next section, the stakeholders were asked for more qualitative perspective on the value of the overall process. Together, these responses intend to validate the following hypothesis proposed in the introduction:

H4: A reasonable number of accepted changes to the system representation will create insights into the domain which were not previously considered by technology strategy decision makers. These insights will inspire a variety of different technology strategy actions.

The interviews conducted in this case study were five in depth discussions with individuals at IBM working in the cloud computing area. The individuals were chosen from different roles in the company, partly to help identify the usefulness of this tool across different roles. In the rest of this chapter, interviewees are referred to by their role. Their roles are described here.

- Consulting Strategist A has the job to help define which business models are most appropriate across a software focused division at IBM, and which technical directions seem to be achieving market momentum.

- Design Strategist A is focused on customer user experience for the same software division, which has many products delivered in the cloud computing model. This includes which features are shown to the end user and how the look and feel of the software is defined

- Product Manager A is focused on defining product requirements for one specific product in what is known as platform as a service, geared towards IT professionals. This involves understanding competitors, interviewing customers, and defining specific requirements for the one product.

- Product Manager B has the same role as Product Manager A, except on an end user focused product delivered via cloud computing.

- Consulting Strategist B has the same role as Consulting Strategic A, except is focused at the corporate level, and is interested in cloud computing business models and technical direction across software, services and hardware systems. This individual was also consulted in the initial construction of the system representation.

Stakeholders were asked to categorize each new link candidate change into one of three categories. The goal of this portion of the interviews was to assess whether the results found were both reasonable to stakeholders in the field, and also resulted in some new and useful insights which made the overall process valuable to the stakeholders.

1. Insights which have face validity and are already known to the expert

2. Insights which are not already known but after further exploration could have face validity.

3. Insights for which the expert is sure they do not have face validity.

In terms of system re-representation, candidate changes that are in category 1 should already be represented in the system representation however may not be in this particular case study because the system representation was constructed with just one of the interviewees. Even in a full execution of this process, items in category 1 may still not be represented in the initial system representation because it is meant to be the combination of perspectives from all stakeholders. Items in category 2 should really form the meat of the system re-representation. These are the links which were not thought of beforehand, but have enough face validity to be worthy of more investigation. Too many items in this category may also lead to an overcomplicated system re-representation, and may require further prioritization of these items to decide what is actually worthy of investigation. Items in category 3 are not to be considered in the system re-representation.

This portion of the interview was conducted in a fast paced manner, but for most interviewees, as they went through the categorization, they paused to discuss their reasoning for the categorization of various topics. These discussions proved to be another means to answering the question of how valuable the overall process was to the stakeholders, and so a summary of that discussion is also included here.

A statistical summary of the results is provided below. Table 5.4 shows the percentage of candidate changes to the system representation that was put into each category by each stakeholder. The three strategy related individuals showed results closer to the expected distribution of results – most candidate changes were known, a good percentage of the remaining candidates had face validity, and a smaller percentage were rejected outright. This is the ideal distribution for the research results. Having most candidate changes already known suggests that the findings from the tool are generally aligned with what experts in the field already believe. Having a reasonable percentage of findings which have face validity suggests that there are indeed useful results to be derived from the tool.

Clearly, the product manager roles exhibited a much more variable set of results than the strategists. The next section on qualitative feedback from the interviews discusses some of the feedback received from the product managers on why this may be the case. Product managers have a more focused role and many of the rejected changes were due to a lack of relevance to their job role. Future study is required to further determine whether this research could be adjusted to be useful to the product manager role or is really meant for the strategist. It is certainly possible that this research is just as valuable to the product managers, if the much more focused set of relevant results has a much higher impact to the product manager's job. The distinction between roles was seen in other case studies and is discussed in the chapter on overall results.

Interviewee	**Known Link (1)**	**Face Validity and Worthy of**	**Reject Outright (3)**

		Investigation (2)	
Design Strategist A	74%	19%	7%
Consulting Strategist A	67%	21%	12%
Consulting Strategist B	72%	18%	10%
Product Manager A	15%	44%	41%
Product Manager B	42%	35%	23%

Table 5.4 Categorization of the New Links.

5.5 Qualitative feedback on candidate changes

After reviewing the specific results, interviewees were asked to provide more qualitative feedback on the overall set of results. Much of this feedback was provided during the actual reviews of the results, as digressions in the conversation. Below, the qualitative results are grouped into the main areas of feedback, and specific examples are provided to help explain the feedback received. The questions posed during this phase of the interview were general in nature – asking about how the research could be useful to the interviewee – and did not specifically ask about any of the below areas unless the conversation led into the specific area.

5.5.1 Overall insight on the cloud computing domain

Overall, several insights about the cloud computing domain were found by this research. These insights were generally driven by the underlying notion that cloud computing is an emerging field which went through a high degree of change in the years studied, 2006 to 2010. Topics were discussed in the blogs before they emerged as generally understood by experts in the field, and examples of this are discussed in section 5.4.2 on time series analysis. Terminology also changes very quickly in the domain, as more popular or generalized versions of old terms take over. This is discussed in section 5.4.7 on term usage.

Many of the insights which fit into the "worthy of investigation" category revolved around the business model subsystem and specifically the pricing options for cloud computing. This was combined with a general notion of cloud computing as a platform for business emerging through the linking of topics such as social networks with topics such as infrastructure as a service. Infrastructure as a service refers to a vendor such as IBM delivering the computing infrastructure (servers, storage, networking, operating system) for customers to run their own applications. The expert interpretation of this phenomenon was that cloud computing moved from a vehicle for specific applications to be available online to a platform for applications to be used together and provide a complete infrastructure for people and businesses to operate online, and this move towards a platform influenced the changes in pricing models from simplistic monthly fees to more complicated pricing models.

The final set of overall insights on the domain is around the audiences for cloud computing. Specific niche markets, such as smaller "mid-market" companies in Russia, emerged as growing areas of interest for cloud computing. In these emerging markets, our stakeholders pointed out that adoption of cloud computing services is viral and when certain companies start to adopt, the speed of adoption leads to other similar companies adopting in the same market. Finally, there was much discussion in the interviews on who was actually blogging in this domain. Are they the users of cloud computing services, or the business decision makers for IT services, or both? The stakeholders' consensus on this question is that the distinction between users and business decision makers is not as relevant with cloud computing because one of the characteristics of cloud computing is that it allows anyone to acquire IT services without requiring an IT infrastructure.

The next few sections detail some examples of more specific feedback received in the interviews.

5.5.2 Time Series

As the field of cloud computing changed significantly in the time period studied, one area of interview feedback was in how there were surprises in when certain links between terms arose in the blogs as compared to when they would have been expected. These surprises are indications that a regular analysis of blogs in the domain may lead to early indicators of new phenomena in the domain, and also that certain phenomena may not appear in the blogs as early as could be expected.

Both Design Strategist A and Consulting Strategist A pointed out that the anchor pair Open Cloud – Social Networks occurred as early as 2006, and was surprised that the cloud term was in use with social networks in 2006. In the group of terms related to this anchor pair, the Design Strategist was surprised that software as a service was only used in 2009. The Design Strategist also noted that the anchor pair On Demand – Social Networks only became associated with time to market in 2010, and would've expected this association earlier.

Product Manager A noted that utility computing stayed popular throughout the time period and would have expected it to drop off over time. They were also surprised that the anchor pair Utility Computing – Social Networks overlapped in timing. Consulting Strategist A also noted this relationship and indicated that the thought of social networking tied to utility computing could help justify a shift in strategy for IT vendors who view social networking as a consumer service.

Consulting Strategist B said that the timing of the link between Time to Market and Upfront Commitment, which emerged during the last few months of the time period studied, correlated with declining economic conditions. This correlation was interesting to the strategist because it indicated that when economic events occur, this research may provide a means for understanding the broader impact to the field.

Given that there were both links which occurred earlier than expected and later than expected, Consulting Strategist A was asked whether this indicated the research results may not be valid.

The reaction was that of course every new link requires investigation to validate, but both types of results could be new insights. Links occurring earlier than expected could indicate early warnings of changes in the field, while links occurring later than expected could indicate that market adoption was not occurring as fast as was thought.

5.5.3 Expected links not occurring

It is important to point out that the links found by this research were not complete or exhaustive. Throughout the interviews, all stakeholders referred to terms which they would have expected to see linked to anchor pairs. Table 5.5 shows a set of sample terms which were expected by Design Strategist A when reviewing the results. The anchor pairs listed were found in the research. The expected terms were expected to be found related to the anchor pairs.

Anchor Pair	**Expected Terms**
installation - subscription	platform, ease of use
perpetual license - cannibalization	registration, payment metrics
utility - utility computing	cost, devices, time to value, privacy, security
open cloud – social networks	cost, pervasive
on demand – social networks	registration, administration

Table 5.5. Sample of terms expected by Design Strategist A

The expected terms comprised of both terms which existed in the initial representation (and thus were used in the research) as well as terms which were not in the initial representation. None of the stakeholders saw a value to distinguishing between these two types of terms. The stakeholders acknowledged that although there were some expected results which were not found, it did not invalidate the value of the results which actually were found. Future study could include a means for testing the relationship between a specific set of terms, to identify the nature of the link between those terms. Future study could also include a means for identifying new terms, through various methods which already exist in the literature and are discussed in the future study chapter.

5.5.4 Justifying known insights

The initial set of hypotheses for this research focused around the possibility of identifying new insights that stakeholders were not originally aware of. However, Consulting Strategist A indicated a number of examples where the results provided data which could have helped justify points that the strategist was trying to propose in their regular work. In the consultant role, suggestions are often provided based on intuition, and this research provides a way for consultants to justify that intuition with more specific data. In the words of Consulting Strategist

A, “you have business people trying to explain the trends as 'headlights' but the decision makers want 'taillights'” and this research provides "important validation for the crystal ball work that I'm doing.” Another potential benefit raised by Consulting Strategist B is in onboarding new members to a team. She said, “every day in every meeting, I'm usually just utilized in teaching others what I know. If this system could be used as a level setting technique, to allow the wisdom of experienced people to share knowledge with inexperienced people, and inexperienced people to share new methods with experienced people, that could be a great use.”

Three specific examples, from interviews with Consulting Strategist A, of known insights justified by data in this research are described below, grouped by anchor pair.

Database as a service – Ecosystem, with Agility as an associated term

The strategist was surprised it came up so early in the time period and would’ve expected it to come up now in 2010. Combinations of database as a service and other offerings are rising rapidly, according to Consulting Strategist A. Requests for Proposals are starting to ask for this and competitors have already put offerings out to the market. When the strategist explained this a year ago, internal discussions rejected the topic. If she could’ve pointed to external discussions, it would have helped justify investment in this area.

Walled Garden - Cannibalization

The strategist interpreted this link as an indication that people are realizing that with the increasing security, walled gardens are desired. If cloud delivered software becomes secure enough through the implementation of walled gardens, then it will cannibalize the sales of traditional software. Vendors of traditional software would benefit from discussions on this topic because it may inform their strategic decisions on when and how to invest in walled gardens.

On Demand - Social Networks, with Time to Market as an associated term

The strategist has been trying to justify the notion that what really matters most to smaller companies is helping them get in touch with their customers by conducting new campaigns through social networks. The impact of social networks on getting a product more quickly adopted in the market (time to market) is something that the strategist could have communicated more clearly with supporting data from the blogs.

In all of these cases, the additional piece of feedback received was that having the source blogs associated with a given anchor pair or new link would be very beneficial in justifying the insights. Simply having the statistical data would not be enough.

5.5.5 New business ideas

As mentioned earlier, the initial set of hypotheses focused on identifying new insights. In the interviews, certain insights were discussed as possible new business ideas for the IT industry. Of course, in each case, further investigation would be required but these examples show the potential for this research to lead to new business ideas emerging.

Design Strategist A discussed the grouping of Walled garden, cannibalization and pilot usage as suggesting a new offering idea to do pilots in a walled garden rather than exposing customers directly to the public cloud. A pilot is when a software offering is setup for use by a company before they decide to purchase the offering. The public cloud refers to software available on a shared basis to everyone, such as Google mail and applications. The typical approach is to conduct such pilots in the public cloud, since it is easy to direct customers to sign up for a trial account of the software without any involvement from the vendor. This new grouping suggests that focusing on a walled garden approach may drive more cannibalization of traditional software, where even the pilot would demonstrate the use of the software in an environment dedicated to the customer.

An example raised by both Consulting Strategist B and Design Strategist A is the link between utility computing, database as a service and agility. Database as a service refers to the ability to offer a database that a software vendor can use to store data without running their own database server. The Design Strategist hadn't thought about database as a service as a means to achieve agility. The Consulting Strategist, once identifying the result, suggested scenarios where agility may be achieved, such as a government scenario where decisions can be made faster based on the availability of broad sets of data.

Three other examples of new business ideas provided by the Consulting Strategist A were the idea of selling hardware on a subscription basis (as is done now for netbooks), changing software to a mutual fund model for pricing (where an initial upfront investment in multiple software packages is made and leveraged over time), and focusing on the Russia small business community as a potential audience for cloud computing sales.

5.5.6 Value of system representation

The interviews were all conducted with stakeholder who did not use system representation in their regular jobs and were not familiar with the CLIOS process. However, certain benefits of using system representation were identified in the interviews. This case study did not conduct any formal experiments to separately identify the benefits of simply introducing the CLIOS process versus the benefits of the overall research methodology including CLIOS and the blog analysis.

Consulting Strategist A described system representation as a way to "surface blind spots" by exposing areas requiring more investigation and focus. Product Manager A used two examples to reinforce this point – the link between regulation and broadband was not a focus area to him beforehand but could affect product offerings which involve a significant use of network resources; and the growing discussion of Sarbanes-Oxley (SOX), a regulation around finances for business and Healthcare Information Privacy Protection Act (HIPPA), a regulation around the privacy protection for health information, within the cloud computing domain, two government regulations affecting the financial and health care industries. He didn't realize industry specific regulation was a major issue with cloud computing, but upon further inspection, he has realized it is a major issue to address when building services which cater to those industries.

System representation also provides a means to "level set and get a quick view of the field" according to Consulting Strategist B. For someone who was not experienced in the domain, a system representation may not be useful, but someone with active experience in the field may be able to scan a system representation and identify key points to investigate.

The final area where system representation may play a role is in suggesting reorganization options for companies. Consulting Strategist A said that the system representation of a domain may provide a way to structure a company intending to address challenges in that domain. "What if we combined a couple of our units into one business unit? What if we end up with three VPs instead of five? Organizations are siloed, and we need data to justify combining organizations," said the strategist.

5.5.7 Term usage

The nature of usage of specific terms in the cloud computing domain was itself a point of feedback from the interviews. Design Strategist A pointed out that the term "open cloud" seemed more popular than had been expected, and the term "on demand" was driven by vendors marketing so it was interesting to see how popular the term actually had become in the blogs. Product Manager A was interested in the use of hardware as a service transitioning to infrastructure as a service, which indicates a broader view of the value of leveraging shared hardware. Product Manager A was surprised to see time to market so prominently displayed in the results, as it is generally a business term. Also, the link between India and broadband was interesting to Product Manager A because the real benefit is the speed of the network (bandwidth) offered by broadband service, but the popular term to use to indicate a faster network is broadband. Consulting Strategist A also noted the use of the term hardware as a service but suggested that the term private cloud should've been used.

Consulting Strategist A said the primary role which could benefit from this set of results is marketing – they could stop using one term and switch over to saying another term if it resonated better. They could create more resonance by picking words which are more frequently used to describe benefits of existing products. The other potential benefit raised by Consulting Strategist A is the ability to test terms with this system to see if people are using the terms that are being considered for marketing campaigns. According to the strategist, this would decrease "human latency" which is the time taken by humans to translate the words used by marketing into the words they use to describe the same concepts.

5.5.8 Requests for enhancements and future study

In reviewing the results, requests were made to enhance the software tool to derive more benefits through further research. These topics are covered in more detail in the future study chapter, but included here since they were brought up in the context of the Cloud Computing interviews. One major request focused on prediction – using the data to predict future links, as well as identifying the right time to react to insights with execution. Another request was to include sentiment analysis, by looking for terms which indicate a positive or negative view of different

topics. The main other request was for a better format to provide the results to stakeholders. The list of results was too detailed, the re-representation diagrams were at too high a level, so further study is required on presentation of the results.

Chapter 6 Kenya broadband case study

6.1 Case study structure and experts interviewed

The topic of this case study is the deployment of broadband technology in Kenya, leading to internet adoption and associated social and economic benefits. Internet adoption in developing nations such as Kenya is very exciting to study because the introduction of supporting technology is done in a more focused and rapid manner than in the US where internet technology evolved over decades. In Kenya, recent investment from the government has led to a fiber cable being deployed between the United Arab Emirates and Kenya, thus providing a significant expansion in the available internet bandwidth to the entire nation. Anticipation of the potential of internet technology, and the recent application of internet technology to many areas, has resulted in a high degree of discussion amongst the public that made studying blogs on this subject particularly interesting. There is also some technical overlap between this case and the Cloud Computing case study because a major application of broadband in Kenya is in being able to leverage cloud computing to participate in the internet without requiring local vendors or businesses to deploya local server and storage infrastructure in Kenya.

This case study is based heavily on a prior research project which leveraged the CLIOS process to develop a complete understanding of the factors which are involved in broadband deployment in Kenya [Omwenga 2006]. One of the main goals of this case study was to examine how the process developed in this dissertation could be applied to a project which has already used the CLIOS process and has a well developed system representation. Omwenga's project interviewed a number of experts – many of whom were interviewed as part of this case study – to develop a system representation for the various factors influencing broadband deployment in Kenya. By studying the nature of blog discussion on the same components which were included in the original system representation, it is possible to see whether any modifications to the system representation can be made. These modifications represent insights which the experts may not have mentioned in their initial interviews when the system representation was developed, or may not have been aware of, or may have emerged after the time of the initial interviews.

This case study also provides an opportunity to focus on a specific region, and a specific goal for that region. The only blogs studied are those that include mention of Kenya, and the initial system representation focuses on a variety of factors all geared to understand the goal of how to drive broadband deployment in Kenya. This case is designed to contrast the Cloud Computing case study, which focused on the growth of a general technology area. Constructing a case with these focus areas provides a means to examine the different applications of the process developed in this dissertation.

The structure of the case study, mapped to the phases of the process, is as follows:

1. Phase 1 – Bibliometric Blog Analysis. This was done using the exact subsystems and components in the original system representation in [Omwenga 2006]. The full component list and a subsystem diagram is presented in section 6.2. Every subsystem was studied, with terms used exactly as they appear in the system representation. The time period studied was January 2006 to June 2010.

2. Phase 2 – Dynamic System Re-representation. The candidate changes to the Kenya broadband system were produced as a list of changes for review by the same experts who were interviewed to develop the initial system representation. Changes produced in this case study focused on new links between components, and new performance measures. The system representation diagram from the initial project was modified to include a sample set of candidate changes, to help demonstrate the concept of re-representation to the experts, however the entire set of candidate changes were presented as a list.

3. Phase 3 – Interviews with experts. As described in Chapter 3, the interviews consist of two sections. First, the entire set of candidate changes is categorized by the interviewee to produce a set of quantitative results on the face validity of the results. In this case study, that portion of the interviews was done by using the researcher who developed the initial CLIOS system representation to act as a proxy for the entire set of stakeholders and review all candidate changes in detail. Second, the experts are interviewed for qualitative feedback on the overall process. In this case study, that portion of the interviews was done with face to face interviews, in Kenya, of many of the experts who were initially consulted. These conversations briefly touched on the specific results found, and focused most of the time on assessing the impact and possible use of this process in the context of each of the experts interviewed.

The experts interviewed for this case study have the following roles:

- Entrepreneur – a founder of the firm Dataposit, which provides data virtualization services to Kenyan businesses interested in hosting websites and data in Kenya. [Mbowa 2010]

- University Professor – a professor of information technology at Multimedia University College who also is Vice Chair of ISACA, an organization which certifies IT professionals to operate in Kenya. [Wallubengo 2010]

- IT Distributor – a sales manager at Westcon, which distributes IT services from other vendors to clients in Africa. The categories of services they distribute include servers and storage, software and security, and infrastructure. [Phares 2010]

- CEO, Kenya Information and Communications Technology (ICT) Board – the director of a government funded non-profit corporation whose goal is to monitor and promote the use of ICT across the country. [Kukubo 2010]

- Lab Director, Google East Africa- a native Kenyan with experience building a Kenya telecommunications company, who was hired by Google to lead their East African office based in Nairobi. [Mucheru 2010]

- Manager, Nokia Africa – a local manager of an engineering team at Nokia who also plays a role in organizing local IT industry events. [Gitahi 2010]

- Permanent Secretary, Information and Communications Technology (ICT), Republic of Kenya – a cabinet level officer in the Kenya government whose role is to oversee the growth and development of ICT in Kenya. [Ndemo 2010]

The experts represent a broad set of perspectives on the impact of broadband expansion in Kenya. In the following section, more detail is provided on specific areas that are being impacted in Kenya which emerged from the interviews.

6.2 Background on Kenya broadband domain

The overall system being studied in this case study is broadband expansion in Kenya, but the driving force for broadband expansion is to have an impact on the social and economic growth of Kenya. Specific areas of impact were identified by the experts, and provide a good context for the results which are provided in the following sections.

6.2.1 Broadband access to individuals who didn't have it before

In 2009, the first fiber cable into East Africa was completed, reaching under water from Dubai to Mombasa. This cable significantly increased the capacity and reliability of broadband services provided into Kenya. The Kenyan Coast Guard protects this cable as a critical infrastructure for Kenya [Ndemo 2010]. With the increasing broadband service, new means for providing access to those who didn't have it before are available. Digital villages, comprised of internet kiosks are being deployed in villages in rural Kenya [Kukubo 2010]. These kiosks are staffed by paid or volunteer consultants who help communicate the potential uses of the Internet to local individuals and help drive usage. Monitoring the success of these villages, and protecting against potential corruption by the consultants, are key concerns of the Kenya ICT Board. One of the predominant uses of the Internet for rural Kenyans is in access to government services such as licensing, and this is being invested in as a means for driving adoption. For businesses, the fiber cable provides the capacity to upload data created in Kenya to sites outside the country. Unlike Europe, where the emphasis has been on keeping data within the country, Kenya is poised to rely on cloud computing to store data externally, because of the capital investment that would be required to store data within the country [Mbowa 2010].

6.2.2 Use of the Internet through mobile devices

Major banks in Kenya have struggled to provide credit to local merchants who do not have the wealth to meet the minimum balance required for personal or small business credit. A new system, known as M-PESA, has emerged as a vehicle for conducting commerce via mobile phones and maintaining a credit history at the same time [Gitahi 2010].

The goal of M-PESA is to allow two individuals to exchange money without cash and without requiring either individual to have a bank account. The system assumes that both individuals have an active cellular phone, and when purchasing the phone, have activated the M-PESA option on their SIM card. The SIM card is associated with the individual, so when an individual seeks to transfer money, they simply send an SMS message to the M-PESA service with the cell phone number of the individual they wish to pay. The M-PESA system receives the transaction request, and transfers money from one individual's account to the others.

This system was designed to avoid the complexities which block individuals from obtaining traditional bank accounts. The main inhibitor is the minimum credit balance required by most banks, which is too high for many Kenyans. The other inhibitor is internet access.The individuals' identities are authenticated using their Kenya national ID number, which means an M-PESA user does not have to have internet access via computer to setup their account. The other goal of M-PESA is to give those who are too poor to have a bank account the ability to still establish a credit history. By using M-PESA, a local merchant or farmer can demonstrate a credit history which allows them to approach larger stores and distributors to sell their goods. This is an example of an innovation which grew in Kenya based on the availability of broadband to conduct the transactions, and the unique needs of the region. Broadband is required to run the M-PESA central service and to connect the M-PESA system to the cell phone vendors and others involved in supplying the service.

Another area where mobile access to the internet is being uniquely applied in Kenya is in the area of basic phone access to social networking services. Nokia is building a specific application which allows the user of a basic phone – one without a keyboard and internet access – to access social networking services via voice commands. [Gitahi 2010]. For Kenyans who don't have access to computers or even smart phones, and also potentially don't have the literacy skills required to use the internet, this provides a means for accessing an emerging area of the internet.

6.2.3 Broadband for Education

As schools in Kenya struggle with distributing and maintaining updated physical books, the Kenyan government is exploring the notion of distributing tablet computers to students, loaded with electronic books and interactive educational content [Ndemo 2010]. This would allow students to access information which was kept up to date automatically via the Internet, and would also provide a means for students struggling with literacy to learn via more visual and interactive means. This particular project also requires innovation around asset management to prevent theft or sale of the tablets.

6.2.4 East African content

A key focus for the Google East Africa lab is to not simply provide tools for the consumption of content and services on the Internet, but also to provide means for adding relevant East African content to the Internet [Mucheru 2010]. When comparing the percentage of the world's population that uses the Internet to the East African population, and then comparing that ratio to the percentage of East African content on the Internet to all content on the Internet, it became clear that without relevant content, it would be challenging to drive Internet adoption in the region. Specific types of content include local music and videos, maps, and wiki based information. This notion that balancing the ability to upload content with the attempt to drive adoption has led to reconsidering the pricing structure for broadband services, and the balance between allocating network bandwidth for upload and download speeds [Ndemo 2010]. While in developed nations, the balance for broadband pricing plans is often towards more download speed in exchange for less upload speed, the ratio is being reconsidered in Kenya because of the desire to drive the creation of local content. As this change occurs, the notion of an "acceptable" broadband speed is also changing [Wallubengo 2010]. This is a particularly interesting

phenomenon for this project because the blogosphere may provide a means for assessing what individuals feel a "fast" internet speed is, and how that notion changes over time.

6.3 Initial CLIOS System Representation

The initial system representation for the Kenya broadband system is taken directly from the CLIOS research project discussed earlier [Omwenga 2006]. In this system representation, six subsystems were studied, as depicted in Figure 6.1.

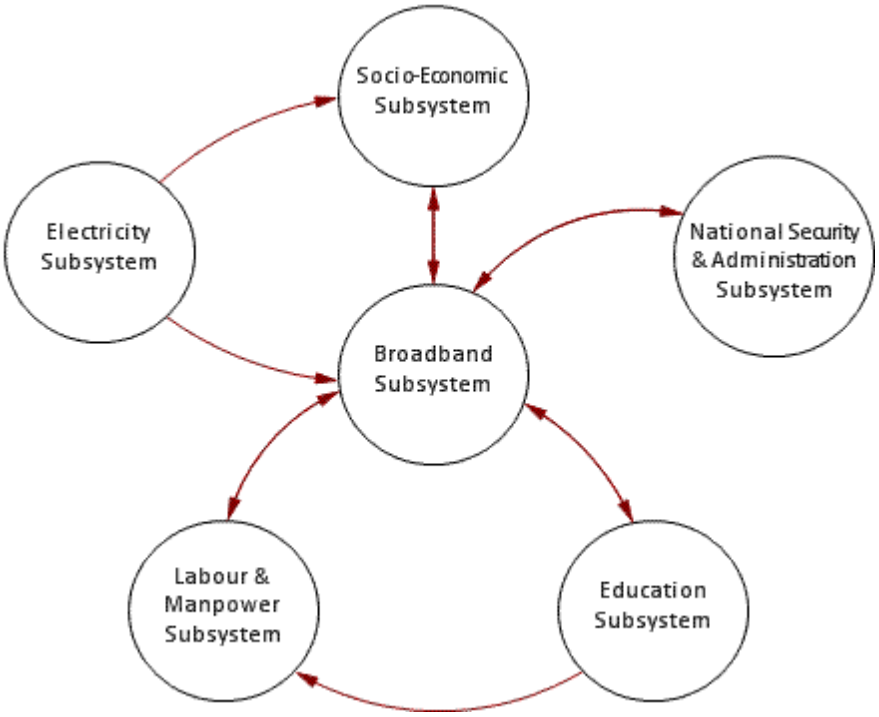

Figure 6.1. Major subsystems of the Kenya broadband system. Source: Omwenga, May 2006.

The broadband subsystem is central to the entire system, and includes the technical components required to deliver broadband services across Kenya. Every one of the five other subsystems interacts directly with this subsystem. The electricity subsystem is included as a separate subsystem because of the significant dependence on electricity which comes with increased dependence on broadband, and the issues around reliability of electricity that exist in Kenya. The socio-economic subsystem includes factors which influence the ability to provide broadband to a broad group of citizens at different socio-economic levels, often through such means as the digital villages discussed in section 6.2.1. This subsystem is related directly to the electricity subsystem because of the correlation between areas that have socio-economic challenges and challenges with reliable electricity. The national security and administration subsystem includes components which cover the government's treatment of broadband as a critical infrastructure, such as the coast guard protection discussed in section 6.2.1. The education subsystem includes both the need to educate citizens on how to use the Internet with the opportunities to further education which exist with increased broadband deployment, such as the tablet innovation

discussed in setion 6.2.3. The labour and manpower subsystem includes the components related to job opportunities which citizens may become aware of and skilled to take up, with access to the Internet. The education subsystem has a link to labour because of the increased skills which can be gained through education received on the Internet.

Each of these subsystems has a set of components which is outlined in Table 6.1.

Labour and Manpower	"industry manpower demand","formal sector opportunities","informal sector opportunities","aggregate labor demand","aggregate labor supply","unskilled labor","opportunity communication labor placement","execution tasks","economic activity","skilled labor","population","e-services products","local foreign investment policy","labor policy"
Education	"institutional learning","certification examination","research development","formal training","skills literacy level","learning mechanism","informal training","basic education","local content services products","elearning","skilled labor","population","certification policy","innovation policy","education investment policy","digital education policy"
Socio-economic Factors	"product process innovation","jobs","production process automation","industry efficiency","collaboration knowledge sharing","transaction distribution costs","income","social service availability","market efficiency","employment opportunities","social impact","market transparency","visibility communication opportunities","economic activity","productivity","manufacture maintenance","innovation policy","labor policy","local foreign investment policy","anti-trust industry policy"
National Security and Administration	"problem occurence","problem identification mechanism","problem communication","problem resolution","resolution mechanism","civic stability","appropriate investment climate","population","e-services products","broadband infrastructure","privacy policy","freedom information policy","national security policy"
Broadband	"broadband connected terminals","ICT equipment imports","aggregate broadband demand","connectivity choice","wire-based connection","wireless-based connection","accessibility","price broadband","GPD per capita","economic activity","macro-economic factors","ICT manufacture maintenance","productivity","broadband infrastructure","electricity distribution grid","e-services products","skilled labor","research development","local content services products","e-learning","e-commerce","e-health","e-security","e-government","BPO","ICT sector investment competition policy","local foreign investment policy","ICT equipment importation policy","tariff regulation policy","e-government policy","BPO policy"
Electricity	"hydroelectricty","geothermal electricty","renewable energy production","non-renewable energy production",“total electricty produced","transmission mechanism","electricity consumption","residential electricity consumption","commercial industrial electricity consumption","aggregate electricity demand","capacity gap","broadband infrastructure","population distribution","population","economic activity","electricity distribution grid","energy policy"

Table 6.1. Component list for each subsystem in the Kenya broadband system. Source: Omwenga 2006.

In Table 6.1, the list of components for each subsystem provides a view of the specific topics which were covered in this study. Each of these components translated into a term which was used in Phase 1 (Bibliometric Blog Analysis) of the study. The reader is referred to [Omwenga 2006] for a complete description of each of the components listed in Table 6.1. Each subsystem

has an associated representation which includes the type of each component and links between the components listed in Table 6.1. For this study, we will select one subsystem to review the representation, as an example that will be carried forth in the results section to show the type of results which were achieved in this study. Figure 6.2 is the system representation for the socio-economic subsystem. This subsystem was selected as the example to review because socio-economic issues were the most prevalent in the interviews conducted with the experts.

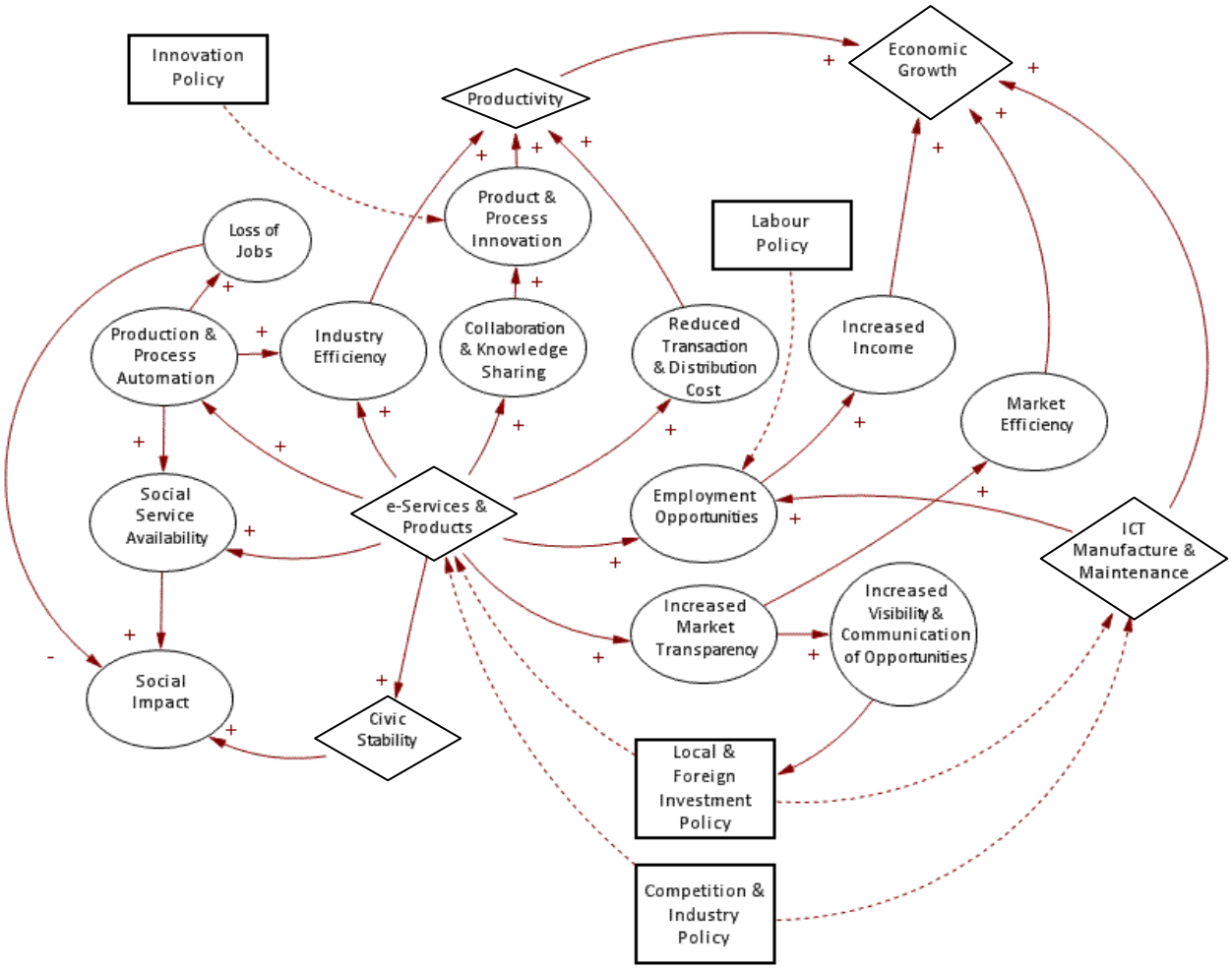

Figure 6.2. Representation of the Socio-economic subsystem of the Kenya broadband system. Source: Omwenga 2006.

In the system representation shown in Figure 6.2, the following component notation is used. Components which are in rectangles are policy levers that are affected by institutions such as the Kenya ICT Board discussed earlier. Components with diamonds are common drivers which means they drive behavior across subsystems and are also replicated on at least one of the subsystems which were shown to interact with this subsystem. All other components are depicted as ovals. The links between components which are a solid line indicate links between components, while the dotted line links indicate the influence of an institution on the subsystem.

The direction of the arrows and the associated "+" or "-" signs in the links represent the positive or negative impact of one component in the other. This notation is not used in every CLIOS representation, but was used in the initial Kenya system representation. For example, the arrow with a "+" sign in the bottom left of the representation indicates that an increase in civic stability would lead to an increase in social impact of broadband technology. However, the link with a "-" sign in the left of the representation indicates that an increase in loss of jobs would lead to a decrease in social impact of broadband technology. For this case study, when new links are identified between components, they are simply proposed as a link, without any indication of direction or positive/negative flow – those details are left to the expert to identify upon further review of the blogs which discuss the topics in question.

6.4 Candidate Changes to the System Representation

The components of the system representation discussed in section 6.3 were put through the bibliometric blog analysis process developed in this dissertation, and a set of candidate changes to the system representation was produced. The cloud computing case study helped refine and fully automate the process for producing these candidate changes, so the changes were all produced in the format shown below, for new links between components. All candidate changes are included in Appendix 2, but a sample is included here for discussion, as well as the modified Socio-economic subsystem representation which includes a depiction of the candidate changes.

The following three "anchor pairs" were identified during the bibliometric blog analysis of the socio-economic subsystem: product process innovation and innovation policy, local foreign investment and innovation policy, and social service availability and social impact. Each of these anchor pairs included a set of components which showed to be related to the anchor pair at various times during the time period studied of 2006 to 2010. The related terms are listed directly underneath each anchor pair. When related terms themselves had a strong relationship, they are listed together on the same line. The month in which the related term peaked in relation to the anchor pair is also listed.

The new links suggested by this analysis are a combination of links between terms in the anchor pair, links between terms in the anchor pair and the related terms, and links between related terms which appear together on the same line. As many of these terms are already linked on the subsystem, and many others were already considered for links, it was relatively quick for the expert interviewee to review these results for face validity and categorize the links into the three categories discussed in Chapter 3. For example, product process innovation and innovation policy are an anchor pair, and thus would be a candidate change to the system representation except that the initial system representation already had a link between them. This is not surprising because the innovation policy decisions made by the government of Kenya are designed to impact the product process innovation which occurs in the domain. However, visibility and communication of opportunities is listed as the first related term under this anchor pair. On the initial system representation, neither of the terms in the anchor pair are linked to visibility and communication of opportunities. This candidate change to the system representation, when examined by the expert, was accepted as a link in the system because one form of innovation policy can be letting people know about opportunities that exist and simply increasing the visibility of opportunities could lead to product innovation. The Permanent Secretary of ICT indicated that when the fiber cable was laid down in 2009, he attempted to

convince the large telecommunications vendors to sell online storage based on cloud computing [Ndemo 2010]. If instead he had made this opportunity more visible to the growing set of entrepreneurial ventures in Kenya, then he may have created a market for online storage.

Socio-economic subsystem bibliometric blog analysis results:

product+process+innovation and innovation+policy - {
anchor pair appears: ['May06', 'Feb07', 'Jun07', 'Jul07', 'Jul07', 'Nov07', 'Dec07', 'Mar08', 'Jul08', 'Jul08', 'Jan09', 'Feb09', 'Feb09', 'May09', 'Nov09'],
"['visibility+communication+opportunities']": ['Feb09'],
"['labour+policy']": ['Jul07'],
"['collaboration+knowledge+sharing']": ['May09']}
"['local+foreign+investment+policy', 'employment+opportunities']": ['Dec07'],
"['income', 'market+transparency', 'industry+efficiency']": ['Feb07'],
"['collaboration+knowledge+sharing', 'labor+policy', 'anti-trust+industry+policy']": ['Jul08'],
"['anti-trust+industry+policy', 'transaction+distribution+costs']": ['Jul07'],
"['economic+activity', 'market+efficiency', 'local+foreign+investment+policy']": ['Nov09'],
"['industry+efficiency', 'labour+policy', 'productivity']": ['May06'],
"['employment+opportunities']": ['Jun07'],
"['employment+opportunities', 'labour+policy']": ['Mar08'],

There are too many candidate changes to include them all on the system representation. However, an illustrative subset was included to demonstrate to the expert the impact of accepting a set of candidate changes. Figure 6.3 shows the same subsystem representation that was shown in Figure 6.2, but with four candidate changes depicted with bold arrows.

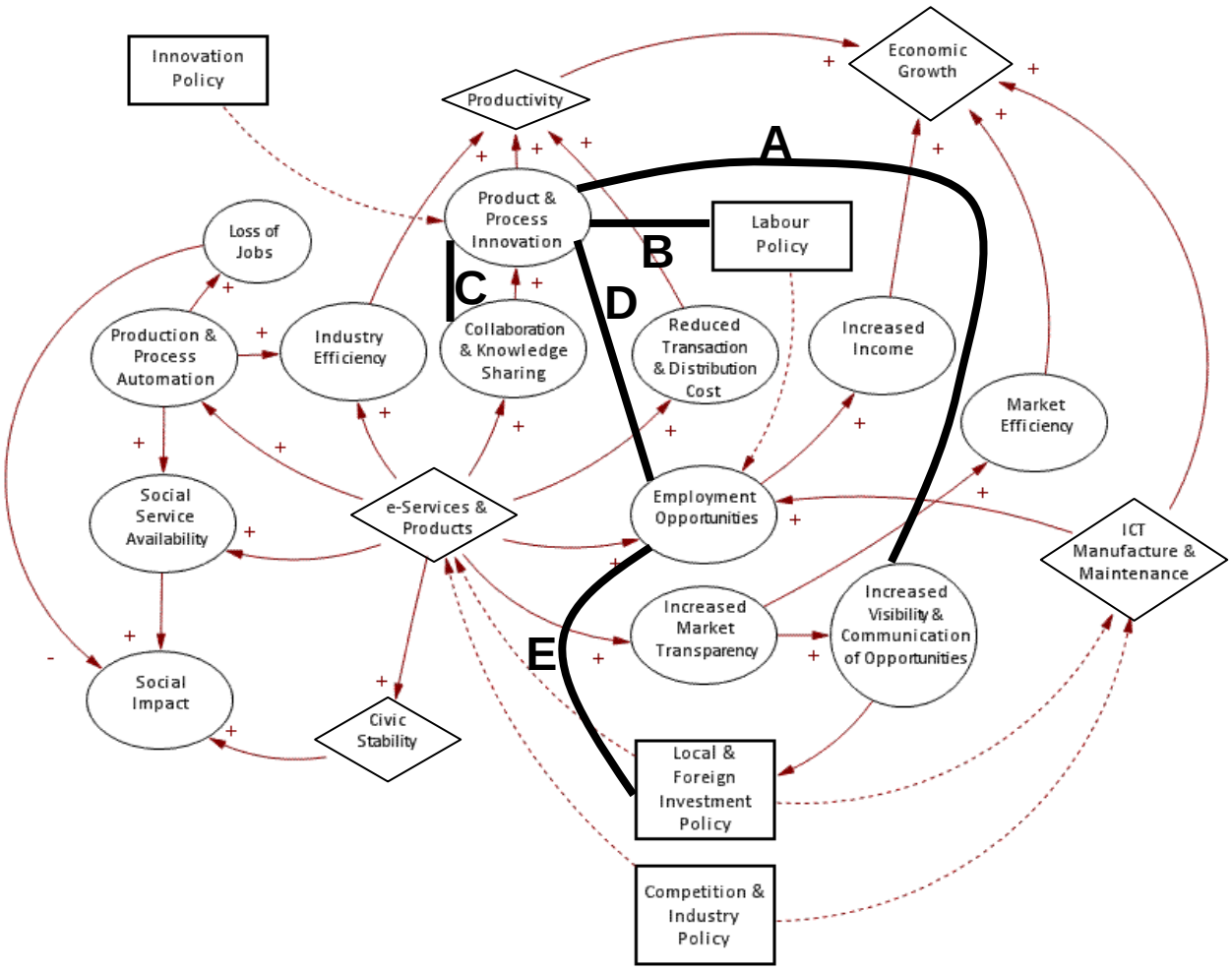

Figure 6.3. Socio-economic subsystem representation with sample candidate changes

The five candidate changes shown in Figure 6.3 are each described here with the feedback provided by the expert interview [Omwenga 2010]. These five are taken from the five related terms which are under the anchor pair "product and process innovation" and "innovation policy". Each of the related terms could be linked to "innovation policy" and to "product and process innovation". In this example, we discuss the potential links to "product and process innovation." No direction is shown in the candidate links because the bibliometric blog analysis does not return any information on the direction of the link.

- Change A: Link between product process innovation and increased visibility and communication of opportunities. This was accepted because, as discussed earlier, making opportunities known more broadly may lead to more innovation occurring. [Omwenga 2010].

- Change B: Link between labor policy and product process innovation. Examples of labour policy are gender-specific programs or retirement age changes. This link was accepted because of the potential to have programs that promote opportunities for certain

parts of the population and the Internet may provide opportunities for segments of the population to showcase their unique skills or expertise [Omwenga 2010]. An example of this would be incentives for women who focus on caring for their family to spend time at the digital villages putting content online that would be interesting to others who are also raising their families.

- Change C: Link between collaboration and knowledge sharing and product and process innovation. This link already exists in the system and when categorized into the three categories for face validity, fits into the first category of known results. [Omwenga 2010].

- Change D: Link between product and process innovation and employment opportunities. This is actually part of a quadruple, since employment opportunities appears on the same line as local and foreign investment policy. Change D was rejected because the expert felt the existing trail of links between these two components was sufficient. [Omwenga 2010].

- Change E: Link between employment opportunities and local and foreign investment policy. This change was described as a "double edged sword" because local and foreign investment policy could have a positive or negative effect on local employment opportunities in Kenya, and the expert would want to further review the nature of discussion around this topic before putting the link on the system representation. [Omwenga 2010]. This could be a case where the link is justified, but the positive or negative indicator is an important aspect.

6.5 Quantitative feedback on candidate changes

The prior section provided samples of the detailed discussion which occurred on the candidate changes. As all of the candidate changes were reviewed, they were categorized into the three categories developed in Chapter 3. As a reminder, these are:

1. Insights which have face validity and are already known to the expert. Face validity refers to an expert's assessment on whether an assertion presented to them has the appearance of being valid [Mosier 1947].

2. Insights which are not already known but after further exploration could have face validity.

3. Insights for which the expert is sure they do not have face validity.

The candidate changes which were produced for every subsystem were put into these three categories, by the expert who conducted the initial CLIOS process and served as a proxy for all of the experts consulted in this case study. These results are shown in Table 6.2

Threshold:	50 months	50 months	50 months	20 mos / < 5 terms	20 mos / < 5 terms	20 mos / < 5 terms
	1: Known	2: Valid	3: Reject	1: Known	2: Valid	3: Reject
Broadband	100%	0%	0%	49%	36%	15%
Education	79%	11%	11%	82%	18%	0%
Electricity	86%	14%	0%	82%	17%	1%
National Security	44%	44%	12%	25%	52%	22%
Socio-economic	n/a	n/a	n/a	55%	45%	0%
Total by threshold	**63%**	**27%**	**10%**	**63%**	**30%**	**7%**
Overall Total	**63%**	**29%**	**7%**			

Table 6.2. Quantitative Results from Kenya Broadband Case Study

Table 6.2 is the summarized set of results from the first phase of interviews which focused on the quantitative categorization of candidate changes. The goal of this portion of the interviews is to assess whether the results found in the bibliometric blog analysis are mostly known, but still produce a large enough set of new insights which have face validity and are worthy of further investigation by the expert. In Table 6.2, each row represents each subsystem. The first three columns, under the "50 months" threshold heading represent the results for candidate changes which were found when the threshold for a change was that it had to occur in 50 or more of the 60 months studied. For example, in the Education subsystem, 79% of the candidate changes that occurred in 50 months or more were already known to the expert, while 11% had face validity and were worthy of more investigation. The last three columns represent the candidate changes which occurred in 20 or more months but occurred in term clusters which had less than 5 terms in them. Splitting the results in this manner provides some insight into whether one type of candidate change produces more useful results than the other.

The main finding from this portion of the results is that the overall categorization of results – shown at the bottom of Table 6.2 - matches the results found in the Cloud Computing domain and represents the intended mix of categories. It is desirable for most results to be already known because it both shows that the system is finding results which generally match the view of the experts and it does not overwhelm the expert with many candidate changes to review. Of the remaining candidates, it is desirable for a larger percentage to be worthy of investigation because

otherwise the expert would spend more time working through candidates that could be rejected outright.

The rest of Table 6.2 shows that most subsystems followed a similar pattern of results, with the national security and administration subsystem being the notable outlier. The expert noted that this was an area which was relatively new to the domain, because the thought of broadband access becoming a consideration for national security only came after adoption grew to a certain level. Thus, interest and discussion in the blogs was likely ahead of experts. For this reason, many of the candidate changes were put into category 2. On one hand, this could represent the most useful part of the finding, because it is the area of least knowledge to the expert, but on the other hand, it could represent an area of "noise" in the blogs where posts are made by individuals with little expertise or serious thought to add to the discussion. Future study is required to help determine whether the ideal use of this methodology is in cases where results are more in line with what was found in the national security subsystem or in what was found in the Cloud Computing case study and in the rest of the Kenya broadband case study.

6.6 Qualitative feedback on candidate changes

The qualitative feedback portion of the interviews is to derive feedback on how this research can have an impact on the decisions made by the experts. The feedback received was similar in structure to that which was received in the Cloud Computing case study. Certain areas of feedback were around the new insights which can be found by studying blogs, and other areas of feedback dealt with opportunities to use the statistical results to justify known insights or monitor changes in term usage or other known areas. The feedback described below was collected via a set of face to face interviews with the experts described in section 6.1.

6.6.1 Changes in Public Perception – Factors, Terms and Issues

Factor Analysis

Multiple experts cited certain factors, such as broadband price and broadband speed, as very critical to the adoption of broadband in Kenya. The notion of acceptable levels for these factors has changed over time according to experts from industry [Mucheru 2010], university [Wallubengo 2010] and government [Ndemo 2010]. In a CLIOS system representation, these factors are often cited as performance measures. But rather than simply identifying new performance measures, the interviewees commented that it would be beneficial to be able to understand what the actual values for these factors are, at different points in time. For example, the acceptable broadband speed will change as Internet users become accustomed to accessing video content from their home [Mucheru 2010].

Changes in Term Usage

The term "bottom of the pyramid" was described by an interviewee as a currently used term for describing the role of developing nations in the world, with large populations but poor economic conditions[Gitahi 2010]. In the IT industry in developing nations, innovations which come from the "bottom of the pyramid" are often discussed using this term. However, while the concept of developing nations having unique needs and unique innovation does not change, the actual

choice of term changes over time, and it would be useful for those who are attempting to follow public opinion on the same topic over time to monitor the changing popularity of terms. Gitahi organizes Open Innovation forums that seek to grow broadband awareness in Kenya and the marketing of these forums and the popularity of topic areas is highly dependent on the terms which are used, since these are often new areas for participants. So selecting the right terms which resonate with participants is critical in the adoption phase of a new technology.

Replacing Surveys

The Kenya ICT board regularly does surveys which look for interest levels on using IT for government services, or other shared services [Kukubo 2010]. After observing some of the results in the socio-economic subsystem, the CEO of the Kenya ICT board suggested that if enough people were participating in the blogosphere, or if this research were adapted to search other sources of public comments, that would provide a vehicle for replacing the currently used method of surveys for garnering public opinion. Kukubo indicated that using this research would allow for latent relationships between topics to be found which individuals may not even directly volunteer when asked themselves. For example, when asked why they are not adopting the Internet, survey respondents may not say directly that they wish there were more employment opportunities to be found on the Internet, but this is a result which did arise from the blog analysis which was done.

6.6.2 Building on Known Insights

Regional Comparisons and Replicating Success in Other Domains

The Indian business process outsourcing (BPO) industry and the means by which IT allowed India to grow employment opportunities for their citizens is one which the Kenya IT community wants to replicate [Wallubengo 2010]. There were likely certain early success factors which contributed to the Indian BPO industry thriving, and analysis of blog posts around the Indian BPO industry may contribute to an understanding of what would make the Kenya broadband industry thrive [Wallubengo 2010].

Justifying Known Insights

This is an area of feedback which was discussed in detail in the Cloud Computing case study. One additional example raised in this case study was in the deployment of the fiber cable. The Permanent Secretary noted that he attempted to tell local telecommunication firms to participate by supporting certain offerings that he felt would resonate best with new consumers of the Internet, but the firms were not willing to listen [Ndemo2010]. If he had a structured set of results as provided in this research, it would do more than traditional public opinion surveys to help convince a business to invest in the areas he knew were the right ones to invest in.

6.6.3 New Opportunities

Identification of Market Demand and Barriers

The entrepreneur interviewed discussed very specific barriers to entry in Kenya, such as the unwillingness of vendors to focus attention on the Kenya market without a good understanding of the market opportunities [Mbowa 2010]. Reviewing the results from this research showed that both market demand can be identified as well as market barriers. Issues such as factors which hinder electricity generation and demand for services such as education are both parts of an end to end solution which vendors can provide via local firms. The distributor noted that technology solutions in the developing world require a different aggregation of technology and thus having a system representation which shows the interconnection of these disparate parts of the solution help a local distributor indicate to vendors which capabilities they need to bundle together to serve the Kenya market [Phares 2010].

Innovation Sharing for Developing Nations

Examples of innovations which are specific to developing nations were raised by the experts – Unilever developed a detergent powder which can be used with river water, M-PESA was discussed earlier as used for microfinance, and tablets for educational content delivery were also discussed earlier. These innovations are often speculated on in the blogs and specific terms which do not appear together in developed nations may appear together in developing nations and indicate a unique innovation [Mucheru 2010].

Monitoring Social Issues

The Kenya ICT board is beginning to see broadband access as a social benefit (not yet a right) for individuals and is starting to monitor broadband access [Kukubo 2010]. This is a combination of monitoring percentages of homes which have broadband access, as is done in other nations such as the U.S., as well as monitoring the style and usage of broadband access. The latter point is potentially more achievable through a systems viewpoint as done in this research as that helps to indicate the potential blockers to providing access to certain parts of the population [Kukubo 2010]. Other issues which affect the adoption and use of broadband in Kenya are intellectual property protection and the immaturity of anti-trust regulation [Wallubengo 2010]. It's unclear to those attempting to promote adoption of broadband, which of these issues could be the most blocking and which specific aspects of these issues are blocking adoption, so close monitoring of the interaction of various elements of the subsystems could also help in this regard.

6.6.4 Expanding Sources Beyond Blogs

A common point raised by many of the experts was that the Kenya IT community focuses a lot of its discussion in email lists rather than blogs [Ndemo 2010][Phares 2010][Kukubo 2010]. The community of IT experts is a generally tight-knit community which has a common shared purpose in seeing broadband expand in Kenya. The permanent secretary of ICT even offered to turn over the entire email list archives for use in this research if the process were adapted to process email lists.

Although email lists may not be the communication vehicle of choice in other domains, this level of interest suggested the possibility that this research has built a generalized process which is not specific to any data source and has value beyond blogs.

6.7 Process adaptation for the Kenya broadband case study

The overall process for this research methodology was largely refined during the Cloud Computing case study. The major adaptation made in this study was in how the interviews were conducted. Since the interviews involved a site visit to Kenya and a single opportunity to meet with each expert, there was not sufficient time to brief each expert on the research methodology and take them through all of the results. This combined with the fact that the CLIOS process had already been executed with the same experts allowed us to use Mr. Omwenga, the original researcher, as a proxy for the experts in doing the quantitative analysis. Mr. Omwenga was also included as a listener in the expert interviews, and this facilitated a more productive discussion because he was able to bridge the domain expertise with the context of this study.

The lesson from this adaptation is that future study should focus on building a process that is able to be absorbed in pieces, for situations where every expert involved does not have the time or interest to learn the CLIOS process or absorb all of the candidate changes. This was taken into consideration when the third case study was conducted, and interview subjects in the third case study were only given a partial set of results to comment on, in exchange for triggering more detailed discussion of the results and the potential impact of this study on their work.

6.8 Kenya Broadband – Chapter Summary

This chapter provided the results of the case study on broadband expansion in Kenya. The domain covered a broad set of experts in various cross functional areas, each with different focus areas for the impact of broadband deployment in Kenya but with a common focus on broadband as a social and economic growth driver. The key insights which were raised were a combination of insights about the domain – such as the focus on visibility as a vehicle for opportunity – and a set of insights about how this methodology could be applied in domains such as this one. Specific insights about the methodology include that certain subsystems may yield unique results because of recent growth of discussion in that domain, and that this methodology could be used to compare specific regions' handling of the same issue. Some of the results discussed in the qualitative results were similar to those found in the Cloud Computing study, such as the changing use of terms and the desire to monitor public perception as two leading uses of this research. Chapter 8 will discuss the common results across case studies in more detail. The next chapter, Chapter 7, is the final case study, on renewable energy. This is similar to the Cloud Computing study in that it is focused on a specific technology area, but is similar to the Kenya broadband study in that it is focused on a specific audience – the professors at the Masdar Institute of Science and Technology.

Chapter 7 Renewable energy case study

7.1 Case Study Structure and Experts Interviewed

This case study focused on renewable energy and academia. Renewable energy is a very broad domain referring to the various efforts to leverage sources of energy that do not face the supply challenges and carbon impact of fossil fuels. The system used for study in this case is an aggregation of two specific technical areas in renewable energy – biofuels and smart grids – and the related aspects of energy consumption and socio-technical factors. These two areas were chosen in part due to the subject matter expertise of the interviewees, because the real goal of this case study was to provide a vehicle for understanding how the research in this dissertation can be applied within the academic community to inform activities such as research funding proposals.

All of the interviewees in this study are members of the faculty or staff at the Masdar Institute of Science and Technology in Abu Dhabi. Masdar is a cross disciplinary institute where the faculty is focused on progressing research around renewable energy solutions from a variety of disciplines. The case study was conducted by spending a week at the Masdar Institute, understanding the various domains in which the professors work, and then conducting interviews in the same format as the cloud computing and Kenya broadband case studies. By spending a week embedded with the faculty at Masdar, it was possible to prepare more focused questions on how the research can apply to the academic community.

The specific fields in which the interviewees work are as follows:

- Biofuels – this professor focuses on the production of fuels based on biological materials. [Hashaikeh 2010]

- Electrical Power and Smart Grids – this professor focuses on the distributed generation of electrical power through "smart grids" which optimize the distribution of electricity [Zeineldin 2010]

- Wave Energy – this professor focuses on harnessing the energy generated from ocean or river waves [Tabaei 2010]

- Materials Science and Solar Energy – this professor focuses on the use of innovative materials to capture and store energy from sunlight [Chiesa 2010]

- Materials Science and Solar Energy – this professor also focuses on the use of innovative materials to capture and store energy from sunlight [Emziane 2010]

- Engineering Systems –this professor focuses on the notion of sustainable energy and the various dynamics involved in large-scale energy systems [Sgouridis 2010]

- Project Management – this individual is a project manager who manages research projects at the Masdar Institute, from funding applications through execution of the project [Prieto 2010]

7.2 Renewable energy domain background

The background on renewable energy was collected by discussions with the interviewees on their areas of research and challenges being faced by each of their areas. This background provided a context for the qualitative interviews which examined the overall relevance of this research to the academic field.

Masdar Institute

The case study was conducted by interviewing professors and staff at the Masdar Institute. This research institute was funded in part by the government of the United Arab Emirates, and founded in conjunction with the Massachusetts Institute of Technology. The Masdar Institute includes professors from a variety of science, engineering, and other fields, whose research all focuses on some facet of the renewable energy domain. The funding backdrop and the location in Abu Dhabi where energy production is not just a technical issue, makes a systems level analysis of the renewable energy domain particularly useful. The Institute is housed within the Masdar City, which is a 6 square kilometer city that is designed to be carbon neutral by implementation of a number of the technologies being developed at the Institute. The political context of energy, the funding and location in Abu Dhabi, as well as the juxtaposition within a city project, makes public opinion and perception a very important input – this arose in the interviews as a significant value provided by this research.

In the following paragraphs, relevant pieces of background on the different areas covered by the interviewees are briefly discussed. Although this case study focused primarily on distributed generation and biofuels as the two technical areas of renewable energy, the broad areas of interest discussed by the interviewees allowed exploration of other subdomains in renewable energy which could benefit from this research.

Solar Energy

The key piece of insight on the field of solar energy is that a system view is needed for success, which includes awareness of funding sources, physical location of the desired solar energy recipient, building design, and materials access [Chiesa 2010]. For example, Masdar City is built on an open desert which has a high degree of sunlight and open space, and can afford to utilize that open space to house solar panels which collect and store solar energy. In Masdar City, the location also presents cooling as a primary use of solar energy, and this use case has influenced the type of solar technology which has been built. Rather than using photovoltaic panels which convert solar energy to electrical power, mirrors are used to collect solar energy and heat liquids that are collocated near buildings that need to be cooled. This is an example of a solution which takes a system view by combining an understanding of the geographic perspective, technical perspective, and funding perspective. The interview feedback discusses the need to collect these perspectives as a motivation for using this research.

Electric Cars

In the oil based culture of Abu Dhabi, electric car adoption has a variety of technical and non-technical factors which influence it [Chiesa 2010]. Some of these factors are unique to Abu Dhabi and others are more generally applicable. Technical factors include the need to support power for air conditioning over long distances. Non-technical factors include the impact of promoting electric cars on the oil-based businesses in Abu Dhabi. This case study did not focus heavily on the adoption of electric cars, but the interview feedback in this area did provide another example of an area of renewable energy which has similar challenges.

Biofuels

In the area of biofuels, the technical focus is on building multi-carbon based compounds as fuel sources [Hashaikeh 2010]. When non-technical factors such as government funding are included, terminology becomes a key factor for which blog discussion may provide useful insights. The public awareness is largely around ethanol, which is a 2-carbon based fuel source, and the general field of biofuels is often equated in public discourse to ethanol [Hashaikeh 2010]. The results section of this case study goes into more detail on the nature of term usage in the blogs, and the similarities between this study and other case studies.

Wave Energy

Wave energy is a relatively new field compared to solar and biofuels, but is trying to learn from the adoption patterns of solar energy and electric cars [Tabaei 2010]. Issues around storage and transport of wave energy, and the geographic proximity of the energy source (i.e., oceans) and the energy consumers, and the type of technology required for oceans versus rivers, are factors that are very similar to what are faced in the solar area. Additionally, the marketing of wave energy and explanation to consumers has the potential to learn from electric cars.

Electrical Power - Distributed Generation

Distributed generation is the technical term that refers to optimizing the distribution of electrical power across a region. This is being achieved through so-called “smart grids” which seek to monitor the utilization of electrical power and then allocate energy appropriately. A major focus of Information Technology (IT) investment in energy is around this area because of the analytics and algorithms required [Zeineldin 2010]. This field, as with the others, also has regional factors – different regions have different local energy distributors and different needs based on how physically distributed the consumers are. Terminology and the use of terms such as smart grids versus distributed generation is also inconsistent between regions [Zeineldin 2010].

Domain Summary

In summary, the renewable energy domain is made up of parallel physical subsystems of technical research, each grounded in a particular set of scientific and engineering disciplines, but with common concerns around energy consumption, regional differences, funding and public perception. As with other CLIOS systems, there is also a set of institutions which affect these physical subsystems, but these were not studied in this case study.

7.3 Initial CLIOS Representation

The goal of this CLIOS representation was to illustrate the potential of the research, rather than to reach any formal conclusions around the domain of renewable energy. The system representation was created by examining content provided by the IBM HorizonWatch program and initial interviews with the Masdar faculty. Given the time constraints around the interviews and the true goal of examining output of the bibliometric blog analysis, the focus was on identifying the right subsystems and components to include in the representation, and thus the representation did not identify a set of initial component types and links and proceeded directly to the system re-representation.

The four subsystems were chosen to provide some analysis of two specific areas of renewable energy, and then two subsystems which cut across all areas of renewable energy. Smart Grid was chosen due to its relation to the IT industry and the prevalence of IT related topics in blogs. Biofuels were chosen because of the prevalence of this area in public discourse and political discussion, which could translate to increased discussion in blogs. Other areas could very well have been chosen, but the needs of this case study to study illustrative examples were met by choosing these two. A socio-economic factors subsystem was studied to include a set of factors which cut across the entire renewable energy domain. The components selected in this subsystem were taken directly from the Kenya broadband CLIOS representation, because those topics represented a good illustrative set of socio-economic factors. The components in the socio-economic factors subsystem were developed outside of this research as part of the initial CLIOS representation done for the Kenya broadband system [Omwenga 2006]. The final subsystem is energy consumption. This subsystem includes a set of components which relate to how any energy source is distributed and consumed by individuals and businesses.

The specific components used in each subsystem are listed in Table 7.1.

Smart Grid "smart grid","average system availability","automated meter reading","building automation system","Conservation voltage regulation","Customer Average Interruption Duration","demand side management","distributed generation","hydroelectric plant","load management","rolling blackout"
Biofuels "bioenergy","biomass","biofuel","cellulose","bioreactor","glucose","enzymes","ethanol","E10","E85","solar energy","moisture","total ash“
Socio-economic Factors (taken from Kenya Broadband) "product process innovation","jobs","production process automation","industry efficiency","collaboration knowledge sharing","transaction distribution costs","income","social service availability","market efficiency","employment opportunities","social impact","market transparency","visibility communication opportunities","economic activity","productivity","manufacture maintenance","innovation policy","labor policy","local foreign investment policy","anti-trust industry policy"
Energy Consumption

"asset management","deregulation","energy service provider","rate of return", "reliability", "residential", "security","service area“

Table 7.1. List of components in the Renewable Energy system.

7.4 Candidate Changes to System Representation

Two forms of candidate changes were included in this case – new links, based on the Latent Semantic Analysis method, and strength of links, based on the Google Similarity Distance method. As with the previous cases, the full set of results is included in Appendix 3, and a sample set is included here to illustrate the types of changes which resulted, and the detailed feedback received on certain changes.

7.4.1 New Links

Sample new links are shown below from the two technical subsystems, smart grid and biofuels. In each, the “anchor pair” is listed in bold, and then the related terms are listed below, followed by comments from the interviewees. Each row has a number following it, which indicates the category which the interviewee placed that specific new link.

The first two sample results are from the Smart Grid subsystem.

hydroelectric plant and distributed generation (9, 8) - 1
{"['load management', 'smart grid', 'demand side management']": ['Feb07'], - 1
"['['demand side management']": ['May07', 'Nov07'], - 1
mart grid', 'load management']": ['Apr07'], - 1
"['smart grid', 'load management', 'demand side management']": ['Jan07'], 1
"['smart grid', 'demand side management', 'building automation system']": ['May06'], 1
'[]': ['Mar06', 'Mar06', 'May06', 'May06', 'Aug06', 'Jan07', 'Feb07', 'Apr07', 'May07', 'May07', 'Jun07', 'Jun07', 'Sep07', 'Oct07', 'Oct07', 'Nov07', 'Nov07', 'Dec07', 'Jan08', 'Apr08'], - 2
"['building automation system', 'demand side management']": ['May06'], - 3
"['smart grid', 'demand side management']": ['Nov07']} – 1
"['smart grid']": ['Feb07'], - 2
In the above set of results, the expert in Smart Grids indicated that the link anchor pair was already known, as hydroelectric plants are used for distributed generation. Of the related terms under this anchor pair, “building automation system’ and “demand side management” were in the third category of being rejected by the interviewee because of the suggested link to hydroelectric plants [Zeineldin 2010]. The term “smart grid” is the one related term which the interviewee put in category 2 (face validity and worthy of further investigation) because although smart grid is the topic area for the overall subsystem, the technical component smart grid referring to an actual smart grid is not always linked to hydroelectric plants.

building automation system and smart grid (4, 1) - 1
{"['distributed generation', 'hydroelectric plant', 'demand side management']": ['May06'], 2
"['average system availability']": ['Aug09', 'Sep09', 'Jan10', 'Mar10', 'May10', 'Jun10'], 2
"['automated meter reading']": ['Jan09', 'Mar09'], 1

'[]': ['May06', 'Jan09', 'Feb09', 'Feb09', 'Mar09', 'Jun09', 'Jun09', 'Aug09', 'Aug09', 'Sep09', 'Sep09', 'Nov09', 'Jan10', 'Jan10', 'Feb10', 'Feb10', 'Mar10', 'Mar10', 'May10', 'May10', 'Jun10', 'Jun10']}, 1

The above anchor pair, “building automation system” and “smart grid”, was also listed in category 1 as a known link. The timing of this link is 2009 to 2010, shown by the months listed in the row with the “[]”. The expert indicated that this timing made sense given the recent introduction of smart grids as part of the considerations for building construction [Zeineldin 2010]. The triple of related terms, 'distributed generation', 'hydroelectric plant', and 'demand side management', were put into category 2 primarily because of the introduction of “demand side management”. Another related term, “average system availability”, was put in category 2 only because the expert expected the timing to line up with the timing of the overall anchor pair.

The following result is from the Biofuels subsystem.

E85 and ethanol (10, 8) – 2
{"['biofuel', 'bioenergy']": ['Feb06'],
"['bioenergy', 'biofuel']": ['Aug07'],
"['bioenergy', 'biofuel', 'biomass']": ['Apr06', 'Sep08'],
"['cellulose', 'biofuel']": ['Oct07', 'Dec09'],
"['biofuel', 'bioenergy', 'biomass']": ['Jun09'],
"['solar+energy', 'bioenergy', 'biofuel']": ['Jun07'],
'[]': ['Feb06', 'Mar06', 'Mar06', 'Apr06', 'Oct06', 'Feb07', 'Apr07', 'May07', 'Jun07', 'Aug07', 'Aug07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Dec07', 'Jan08', 'Jan08', 'Feb08', 'Feb08', 'Mar08', 'Mar08', 'Apr08', 'May08', 'Jun08', 'Jul08', 'Jul08', 'Sep08', 'Nov08', 'Nov08', 'Dec08', 'Jan09', 'Jun09', 'Jul09', 'Jul09', 'Aug09', 'Sep09', 'Nov09', 'Dec09', 'Dec09', 'Jan10', 'Mar10', 'Apr10', 'May10', 'May10'],
The main feedback from the expert on this set of results is that ethanol was part of the anchor pair, and terms such as biofuel are related terms that only appear in certain months [Hashaikeh 2010]. Ethanol is a two-carbon based compound, and one specific form of biofuel. Although much of the research innovation is in multi-carbon based componts, ethanol is a much more well-known term publicly, and the expert would like to investigate further to understand whether research proposals which seek to study biofuels would be better served to include the word “ethanol” if going to government bodies or other agencies which are influenced by public perception.

The following anchor pair is a sample result from the socio-economic factors subsystem.

visibility+communication+opportunities and production+process+automation (13, 3)

This link was put in category 2 because the expert observed that production process automation in renewable energy depend on visibility and communication. Many of the technologies in renewable energy are new and production processes have not yet been optimized. With communication of opportunities, more opportunities for automation can be identified [Emziane 2010].

7.4.2 Common Drivers

The next set of results collected in this case study were around identification of new common drivers. A common driver is a component which appears on multiple physical subsystems and provides a means for changes in one subsystem to help drive changes in another. These drivers were identified by comparing the Google Similarity Distance of each pair of terms between each subsystem, for each month. The goal was to identify those pairs of components across two subsystems which had a growing correlation to each other, because this could indicate that a common driver may exist which links the two subsystems. So, for the four subsystems, 6 tables were produced, one for each pair of subsystems. In each table, the term pairs which grew fastest in similarity over time were proposed as indicators of common drivers. In a system re-representation where a pair of terms across two subsystems is identified as being linked, this could either be represented by replicating a component from one subsystem to another, or by declaring a new independent component as a common driver that appears on both subsystems and is linked to each pair.

Table 7.2 is the table of fastest growing pairs from the Smart Grid vs Socio-economic Factors.

Cat.	Smart Grid	Socio-economic Factors
3	rolling+blackout	market+transparency
1	Conservation+voltage+regulation	industry+efficiency
1	Customer+Average+Interruption+Duration	industry+efficiency
1	hydroelectric+plant	jobs
3	smart+grid	collaboration+knowledge+sharing
3	smart+grid	social+service+availability
2	rolling+blackout	industry+efficiency
3	smart+grid	production+process+automation
3	building+automation+system	social+service+availability
3	building+automation+system	income
1	smart+grid	employment+opportunities
3	building+automation+system	production+process+automation
3	building+automation+system	transaction+distribution+costs
3	average+system+availability	transaction+distribution+costs
2	automated+meter+reading	transaction+distribution+costs
1	building+automation+system	manufacture+maintenance
2	automated+meter+reading	social+service+availability
3	building+automation+system	economic+activity
3	building+automation+system	employment+opportunities
2	smart+grid	income
3	smart+grid	manufacture+maintenance
2	automated+meter+reading	anti-trust+industry+policy

2	smart+grid	anti-trust+industry+policy
2	building+automation+system	labor+policy

Table 7.2. Candidate common drivers between the Smart Grid and Socio-economic Factors subsystems

In Table 7.2, the pairs listed each have one term from the Smart Grid subsystem and one term from the Socio-economic Factors subsystem. The interviewee was asked to place each pair in each of the three categories which are being used in all of the quantitative interviews [Emziane 2010]. For this type of candidate change, category 1 implies that the pair is represented in some fashion by an already known link between the subsystems. In other words, each pair in category 1 is not itself tied to an independent common driver. Pairs in category 2 are the ones which are worthy of further investigation, and which the interviewee believes there are potential new common drivers to be identified in the manner described earlier.

Three pairs from Table 7.2 which all fell into category 2 are discussed here. The components "rolling blackout" and "industry efficiency" might indicate that decisions made in the technical subsystem around how rolling blackouts are enacted could drive industry efficiency [Emziane 2010]. In this case, it may be prudent to place the "industry efficiency" component on the Smart Grid subsystem, and identify it as a common driver between the two subsystems. The components "automated meter reading" and "transaction distribution costs" are worthy of more investigation because it may indicate that the generic notion of "automation" is a common driver between the two subsystems [Emziane 2010]. In this case, it may be prudent to create a new component, "automation" and place it on both subsystems and then link each of the components "automated meter reading" and "transaction distribution costs" to "automation". Finally, "smart grid" and "income" are put in category 2 simply because the relationship is not clear to the interviewee but if there were a relationship, it would be of interest [Emziane 2010].

7.5 Quantitative feedback on candidate changes

Six of the experts were each asked to go through the candidate changes and place changes into each of the three categories which have been used in this research. Due to time constraints in the interviews and the desire to have sufficient coverage of the qualitative portion of the interviews, each interviewee did not cover every candidate change. Interviewees with specific subject matter expertise in specific areas started the interviews with their own areas of expertise.

Interviewee Area of Expertise – Energy Case Study	Known Link (1)	Face Validity and Worthy of Investigation (2)	Reject Outright (3)
Biofuels	100%	0%	0%
Project Manager	31%	69%	0%
Engineering Systems	68%	23%	9%
Wave Energy	24%	28%	48%
Materials Science/Solar	92%	8%	0%
Electrical Power	79%	17%	4%

Table 7.3 Categorization of candidate changes in the renewable energy domain

Table 7.3 covers the overall categorization of candidate results by the various experts interviewed. This is similar to the table produced in the Cloud Computing study, where each row represents the categorization for a distinct expert. The main result shown in Table 7.3 is that each interviewee had a different spread of results between the categories. The biofuels expert put all of the candidate changes into category 1, whereas the project manager and the wave energy expert put less than a third of the changes they reviewed into category 1. The biofuels expert only reviewed results in the biofuels subsystem, and indicated that while terminology choice was worthy of investigation, the underlying links between components were all as expected and it was hard to distinguish when a link in the subsystem was warranted because the goal of the subsystem was not clear [Hashaikeh 2010]. The project manager indicated that he placed many of the links in the worthy of investigation category because he did not have the technical expertise to identify many links as category 1 [Prieto 2010]. The wave energy expert indicated that he placed many links in category 3 which were not relevant to his field, as his field is new and focused on technical growth as opposed to the various socio-economic and consumption factors in other subsystems [Tabaei 2010]. The engineering systems, materials science and electrical power experts all fell closer to the expected distribution of results across the categories.

In summary, these quantitative results reinforce the notion that different roles have different perspectives on the utility of this research. It is good that none of the interviewees had a blanket rejection of the research, and all of them found some candidate changes which were worthy of investigation. Half of the interviewees responded with categorization that was in the expected distribution. The qualitative interviews demonstrated that every interviewee had perspectives to share on how they would use this research.

7.6 Qualitative feedback on candidate changes

The following sections describe the qualitative feedback received from the academic experts interviewed in the renewable energy case study. These elements of feedback have some overlap with feedback received in the prior case studies, and Chapter 8 provides full coverage of the areas of consistent feedback between the case studies.

7.6.1 Overall insight on renewable energy domain

The interviews highlighted the fact that each of the different subfields in renewable energy all have a common set of goals around driving adoption of their technology and effective distribution of energy. Thus, even though the technology choices are different, technical patterns can translate from one field to another, and non-technical patterns almost always will translate [Hashaikeh 2010].

The field of renewable energy is also a heavily political field and thus has a number of non-technical actors whose views on technical choices often have funding consequences. Reconciling term usage is one way in which communication between technical and non-technical actors can be bridged [Prieto 2010].

The final observation is that different fields are at different stages of the maturity lifecycle in terms of adoption. For example, wave energy was cited as being very early in the lifecycle as compared to solar energy. However, it is certainly possible for fields to learn from each other as they progress. For example, early investment in electric and hybrid cars has been in small vehicles rather than large SUVs, primarily due to feasibility of powering a large vehicle with electric power. The field may learn from whether adoption is best driven from this market or perhaps could benefit from focusing on SUVs where gasoline usage is highest. Wave energy is similarly focused on oceans, due to feasibility, but may find that looking at rivers may have more impact on consumers [Tabaei 2010].

7.6.2 Term Usage

The use of certain terms and the relative popularity of certain terms came up as a key insight which this research could be used to obtain. For example, in biofuels, the prevalence of the term ethanol as a proxy for the overall field of biofuels was obvious from the results [Hashaikeh 2010]. The linkages between the term ethanol and other terms in the subsystem provides a means for understanding whether researchers should use ethanol as an explanatory term when presenting to non-technical individuals.

In other fields, the same term may have different meanings, and seeing the terms it is linked to help explain the various uses of the same term. This is especially useful in new and changing fields such as renewable energy. In wave energy, the term "security" can have different meanings, it could refer to the security of the energy production (wave energy plants are often on islands), or it could refer to the security of the distribution network [Tabaei 2010]. In distributed generation, the scale factor of terms is an issue, similar to what was reported in the Kenya broadband study. A "small" hydroelectric plant can have a different meaning depending on the region, or can change over time as capabilities increase [Zeineldin 2010].

Finally, understanding term replacement was mentioned by multiple interviewees as a possible use of this research. As the field changes, certain terms will become more popular the others for referring to the same concept. For example, the term biofuel may indeed replace ethanol once public understanding and perception matches scientific understanding. Understanding the

changing use of terms, and possibly even mapping it to the use of the same terms in technical papers is another possible application of this research.

7.6.3 Preparing for funding and regulation by understanding public opinion

One of the main reasons that the researchers found interest in understanding the public's use of terms is because the funding sources they utilize are often more in line with the public's understanding of the field. In the ethanol example, funding proposals may benefit from including or avoiding the use of the term "ethanol" depending on public perception of that concept [Hashaikeh 2010]. The same technical research could be proposed with a set of terms that has the most positive connotation at the time. This is very similar to the feedback received in the cloud computing case study around finding the right terms to market the same technology.

Another challenge faced by researchers is the need to explain their work to non-technical audiences, either in the form of introduction sections of papers, or at talks given to more general audiences. With a changing field, even when researchers attend conferences in their own discipline, the topics related to renewable energy may not be familiar and an understanding of public opinion could help attract a broader audience [Tabaei 2010].

There is also a phase delay in public understanding versus technical progress [Emziane 2010]. Often times it takes months or years before the public is familiar and comfortable with topics which have been identified by scientists. This is an interesting result because it is directly opposite the finding in the cloud computing study that the public interest in a topic was seen ahead of when the industry experts delivered products to meet the interest. Both sides of this notion are discussed in the summary of results provided in Chapter 8. Understanding the duration of this phase delay could also help focus effort on shortening it. As stated by one of the researchers, "we live in society, and as scientists, researchers, engineers, we have more duty to contribute to the better understanding of the reality of things" [Emziane 2010].

7.6.4 Regional Differences

Regional differences in terms of the requirements for renewable energy as well as the perception of different elements of the field are another potential insight to be gained from this research. The project manager who prepares funding proposals cites the need to observe how certain regions view certain technologies. This study does not explicitly study blogs in a particular region, however the Kenya broadband study did focus on blogs which mention Kenya. Future study could scope blogs to particular regions to study differences between regions. In cases where a major event such as an oil spill adversely affects a region, it may cause a regional difference in how certain technology is perceived in that region [Prieto 2010].

Distributed generation is a term which has synonyms or related terms in various regions, such as "embedded generation", "distributed resources", and "dispersed generation" [Zeineldin 2010]. This is true even though there are IEEE standards for electrical power distribution which mention certain specific terms. There is a desire to capture all of the regional differences as well as track the changing use of terms as regional differences decrease over time.

Regional differences also impact the requirements for certain forms of energy. For example, solar energy is highly dependent on the availability of sunlight in the region, the temperature, and physical space available for solar energy collection and storage materials [Chiesa 2010]. Regional consumption of energy is also different, as certain regions are more urban,and may have more load at certain times. These differences also translate to differences in funding availability and investment.

7.6.5 Filtering Blogs to focus on what to read

When researchers were asked if they already consult blogs as an information source in their regular work, the main comment raised by each was the desire to filter blogs and prepare a subset that were worth actually reading. Even if certain blogs written by respected experts are already selected, it would be good to use this research to identify new blog topics or authors who may be worthy of reading [Sgouridis 2010]. One method of selecting blogs which are likely written by experts is to search for blogs which contain a sufficient number of technical terms. This method only works if the technical terms are not well known by the public and included in all blog posts in the area. One of the benefits of renewable energy being a new field is that the general public has not yet developed a familiarity with all of the technical terms involved [Emziane 2010].

7.6.6 Deeper understanding of technical areas

Understanding common patterns at a high level

The topics covered in this case study were generally at a high level and provided some semblance of patterns between different technical areas. For example, if biofuels and smart grids are considered peer areas, this case study provides some insight on the general discussion of each field and can help identify common patterns such as public understanding of the field or adoption factors related to the field [Hashaikeh 2010]. It would be more effective, according to two interviewees, to do this level of analysis of a broader set of approximately 10 peer areas, and then use the results to both show how learnings from one field can be applied to the other, as well as how scientific work in one area may be explained in the context of another [Tabaei 2010] [Hashaikeh 2010]. For example, carbon capture is a major area of scientific progress related to biofuels, but if the solar energy field demonstrates an interest, then the same research may be applied to solar cells which capture sunlight.

Understanding deeper technical insights at a low level

Interviewees consistently requested deeper technical coverage of their individual domains as a follow up analysis to the general topics covered in this case study. In certain cases, interviewees suggested drilling down on the specific blogs or specific topics around areas which seemed counterintuitive but still had some face validity [Chiesa 2010]. For example, if a link is suggested which could be plausible but does not seem obvious, then doing an analysis of deeper technical terms in the same area may help make the decision on whether to accept it. One direction to consider here is to use term discovery, which is described in the future study section of Chapter 9. Term discovery allows new terms which are related to an existing term to be

identified without a human expert prescribing which terms to search. This would allow the many terms which are related to a general topic such as biofuels to be studied.

7.6.7 Growing beyond technical research

Transition from R&D to Commercial Products

All of the interviewees in this case study were in academia but many have an interest in seeing their work translated from research and development to commercialization. To accomplish this, the evolution of certain fields which have seen market success may influence the choices made by other fields. Understanding past behavior may provide a means for predicting the future behavior of another field and driving adoption or regulation [Tabaei 2010]. Further work on this case study could include interviews with researchers and managers from industry who are working on similar areas of renewable energy. Experts from industry may have insights into adoption and product development which could inform the academic researchers interviewed in this study.

Common Drivers between technical and non-technical subsystems

One of the ways to best understand what factors drive success of a certain technical area such as biofuels is to develop a better understanding of common drivers, or components which drive behavior from the technical systems to the non-technical systems. In the field of biofuels, it is interesting to see what factors drive connections between technical aspects to topics such as job creation [Hashaikeh 2010]. For example, blogs in the area of renewable energy which mention job creation also mention hydroelectric plant. This may indicate a public perception that physical plants have the need for employees and may drive job creation. The reality may in fact be that other areas of technical investment have a greater impact on job creation. However, in this example, the perception of what drives job creation may be just as useful as the actual factors which drive job creation, because it helps focus communication to the public as well as regulators who may be influenced by the public.

7.7 Process adaptation for the renewable energy case study

In this case, the main process change which was made from prior cases was the simplification in generating the initial CLIOS system representation. Instead of producing a full system representation, the focus was on which terms would be sufficient to illustrate the use of this research. This process adaptation also raised the potential of using this research to do the initial generation of the CLIOS representation for a system. Future study could pursue this further, but would have to include some form of term generation to generate terms which are not initially identified. The other major process point raised in this case study is the time spent understanding the work background of each professor to be interviewed. The system studied in this case was much broader and had a less focused goal than the other two systems. Spending time with the interview subjects allowed the formal interviews to be more focused around potential uses of the research to their work. This suggests that the actual deployment of this research in a real world setting would require a system analyst or consultant to assist in making it useful to stakeholders in a given domain.

7.8 Chapter Summary – Renewable Energy case study

This chapter provided the background and results for the renewable energy case study. This case had the unique aspect of focusing entirely on an academic audience and on the broad area of renewable energy. The actual results of the bibliometric blog analysis were in line with those produced in the other two case studies. The results from the quantitative interviews showed that certain interviewees categorized the results in the expected distribution – with most results marked as already known and a good number of results having face validity – but other interviewees categorized the results in a much different distribution. The feedback provided in the qualitative portion of the interviews had some overlap with the other two case studies – term usage, regional differences, and using this research to justify known insights were examples of overlapping feedback. This case study also produced some unique feedback in the ability of this research to potentially map progress in certain fields to other fields, and in the ability to identify common drivers between technical and non-technical subsystems. Chapter 8 takes the feedback received in all three case studies and aggregates it into a consistent set of feedback which helps examine how this research has performed against the hypotheses raised in the introduction of this dissertation.

Chapter 8 Overall Results

This chapter covers the major themes which resulted from the case studies. Together, these themes represent an overall set of feedback received on this research from the three case studies. The goal of this chapter is to provide perspective on how this research can be applied in practice. Many of these areas naturally motivate future study to further explore these areas, and in these cases the opportunity for future study is mentioned briefly and covered in more detail in Chapter 9.

8.1 Major Themes

8.1.1 Support Intuition with Analytics, based on Fact and Perception

All three case studies included feedback that the bibliometric blog analysis provided a set of analytics which helped justified intuition that the experts already had. While the initial goals of this research were to help unveil new insights, it became clear that even justifying known insights is itself a useful result. The experts in a particular domain may have a good understanding of their field which allows them to make intuitive decisions, but they still benefit from having explicit data to show others when trying to justify their intuition. In the cloud computing study, the strategists pointed out that they had identified market phenomena such as the proliferation of social networking in the enterprise much earlier than it was acted upon by their product development counterparts. With proper data to identify both the proliferation of enterprise social networking and the detailed factors which affect adoption, cost and other goals, a strategist could have put forward a much more concrete case for investment. Since blogs are based on the author's perception, the results of this methodology represent one collective view of the public perception in the field. An expert can examine whether public perception matches their view, and use this knowledge to emphasize, reword, or change the way they articulate what they are proposing.

8.1.2 Insights Can Be Used to Bring Together a Variety of Roles and Regions

Interview feedback showed that in applying this research, it is important to consider the specific roles of the experts being interviewed, and the specific geographic regions being considered. In each case study, people with different roles were interviewed, and interview feedback was different among the people with different roles. In some cases, differences in roles represent different levels of knowledge on specific topics, different priorities, and other differences which may influence the expert's use of the results. The information provided by this research acts as a common denominator. For example, the product managers interviewed in the cloud computing study did not have the breadth of knowledge available to the strategists and used this research to gain a broader perspective and exchange ideas with the strategists. In the Kenya case study, the insights on adoption helped bring together government experts seeking to promote adoption with industry experts seeking market opportunities.

Similar to bringing together different roles, interview feedback also suggested that this research could bring together different regions. Regional differences in terminology or articulated

concerns can be exposed by studying blogs across different regions. Currently, this research does not distinguish by region. However, interviewees stated that they still saw significant regional differences in both the use of certain terms as well as the perception of different factors such as "broadband speed." Similar differences were also seen in the Energy case study, as discussed in Chapter 7.

Future study is required to explore both of these dimensions further. Extensions to this research can look at the different methods of blog search which were discussed in Chapter 2. These may require an index of the blog data to study meta-data or may require enhancing the search algorithms which was used in this research. Blogs are unstructured information but they do have some meta-data which can be utilized – the blog author, sometimes a profile which includes location, and often a blog roll which provides some additional context on who the blogger is. This requires a different methodology for collecting information from blogs – two options discussed in Chapter 2 are advanced search and indexing of blog data.

8.1.3 Domain and Subdomain specific differences

The three domains which were studied were each different in scope and in adoption lifecycle. In terms of scope, the cloud computing and renewable energy domains were broad coverage of a field, where the Kenya broadband domain was more focused around a particular region's adoption of a technology. In terms of adoption lifecycle, cloud computing was relatively further along in growth and maturity than broadband adoption in Kenya, and the renewable energy domain was made up of subdomains which were each at different stages of the lifecycle. These differences, combined with the rate of change of each of these domains, demonstrated that this research can be used across a broad spectrum but requires explicit identification of the scope, the stage of adoption, and the degree of change. Identifying these factors upfront helps drive the interview topics discussed with the experts and helps focus the discussion of potential uses of this research.

Interview feedback suggested that properly identifying these factors would also lead to being able to map insights from one domain into another, and one subdomain into another. Certain patterns often emerge, especially around the linkages between technical components of the system and social or economic factors. For example, since cloud computing is further along than broadband adoption in Kenya, the link between social networking and the enterprise which was found in the cloud computing study may help predict what will occur in Kenya as broadband adoption increases. Although this may be an intuitive prediction to an expert in the field, the specific relationships between terms in the cloud computing study provides a fine grained view towards what specific factors could lead to the link occurring in Kenya.

Identification of scope of a domain or subdomain can be done automatically in future study, perhaps by looking at co-occurrence of terms or including keywords which designate different levels of scope. For example, the study on broadband adoption in Kenya could be re-done to be scoped to just Nairobi, or scoped to all of East Africa or all developing nations. In the Nairobi case, a keyword such as "Nairobi" could be included in the searches to collect blogs which are concerned with broadband in Nairobi. In terms of adoption lifecycle and identifying patterns, future study can explore whether the same terms be found over time, with regard to one topic in

one year, and then with regard to another topic in two years. This would allow a more systematic means for transferring insights from one subdomain to another and contribute to organizational learning.

8.1.4 Expected Links Not Occurring

In certain cases, interviewees indicated that links they expected to see come out of the analysis were not occurring. There are four possible categories for these expected links

1. The blogs do not contain complete information. This means the blogs do not discuss all topics and relationships known in the field, especially in cases where experts in the field simply do not write blogs.

2. The algorithms and parameter values selected do not appropriately characterize the information within the blogs. This would indicate that a human reader of the blogs would notice a relationship between two terms, but the particular choice of algorithms made in this research would not reveal that same relationship.

3. The expected links are so obvious to blog authors that they are not discussed explicitly . This category is for those links such as "sunlight" and "solar" which are so obvious that they are not mentioned together in blogs.

4. The terms which are expected to be linked are actually not considered linked by the blogging community. This category is for those terms which an expert may consider to be related but a general member of the blogging community, if asked, would actually disagree that there is a relationship, and thus would not blog about it.

Further study is required to appropriately split results into these categories.

For those links which fit into the third or fourth case, there may be useful insights to be derived from a lack of discussion in a certain area. If a link is considered intuitively obvious and thus not worthy of explicit discussion, it may indicate progress in terms of adoption. If links were expected but actually not considered linked by the blog authors, this may be a sign that the experts being interviewed have not appropriately understood that a field has moved on from a certain topic. In the renewable energy field, for example, technical experts may believe there is public interest in a certain technology but may not realize that interest has in fact dissipated over time.

8.1.5 Time series is relevant both in the past and future, especially for factor analysis

The initial goals of this research were centered on technology forecasting and how analysis could predict future trends for the purpose of developing a technology strategy. However, experts expressed significant interest in looking at time series data in the past to understand significant historical events. For example, in cloud computing, the strategists were interested in the point at which subscription pricing became prevalent. In renewable energy, one expert was interested in how the use of terms changed over time, for example, the popularity of the term "smart grid"

versus “distributed generation”. In Kenya broadband, the meaning of the term “fast broadband” was of interest. In each of these cases, observing time series data in the past was seen not just as a means for predicting the future, but as a relevant insight on its own, especially when mapping past events to decisions made by the experts.

One key concept emerging from this area of feedback is the ability to assess a phase delay in discussion between experts and the public, as opposed to the reverse. The initial goals of technology forecasting assumed that blogs would contain ideas which had not yet been included in expert discussion found in journals and conference proceedings. While this is true, there are also certain insights which are known to experts but then experience a delay in getting discussed in the blogs. Blogs are written by individuals with varying levels of expertise, so they provide an approximation of public discussion which can be compared against pure expert discussion in other venues. Understanding both the duration and the cause of this delay could help experts in shortening the delay in the future. On an informal level, the experts interviewed in this study were able to do this because they were given the dates when certain links started to occur and could correlate those dates with their own view on when the links were known. Examples of this, such as the timing of the link between enterprises and social networks, were discussed in the three case study chapters. Doing this in a more formal manner would require structured analysis of the expert views (perhaps by applying this research to journal articles or conference proceedings) to compare to the results from blogs.

8.1.6 Exposing new ideas

Perhaps the most powerful result to come out of the case studies was in situations where new business or research ideas were identified. Examples of these new ideas were found in each case study. In cloud computing, new business models for offering hardware as a service were new to the expert strategists. In the Kenya case, the link between visibility of opportunities and economic success could lead to new ideas for the government experts. In the energy case, the importance of studying adoption of renewable energy technology can lead to new research projects around adoption.

In reviewing the examples of new ideas which emerged from the three case studies, it is clear that the use of a system-based analysis method such as the CLIOS process is vital in exposing new ideas. Without the context of a system representation, or the structure to look for characteristics such as performance measures, the bibliometric blog analysis would be focused simply on popularity of different terms, and the relative popularity of a term would not itself lead to a concrete new idea such as the examples described above. The three case studies in this research did not explore whether the introduction of the CLIOS process was itself what generated the new ideas.

8.1.7 Informing the Execution of Ideas

Another benefit which emerged from taking a systems approach with the CLIOS process is that the blog analysis helped experts turn their ideas into action. Examples of action which could be taken on expert ideas include getting ideas to market, submitting a proposal, or promoting a piece of regulation. In all of these cases, the blog analysis provided insight not just into the

actual idea, but also into the considerations involved in the mechanics of executing the idea. For example, in the Kenya case study, the idea of using broadband to provide educational content in schools requires an understanding of the need to monitor devices shared with students and the literacy challenges which may prevent a student from benefiting from the educational content. Communication is another key element of execution – in taking an idea to market, explaining the idea to the public in a way they understand is critical, and the analysis provided in this research helps inform the best way to communicate.

8.1.8 Benefit is more around value than cost savings

A valid question to consider in evaluating the application of this research is in how to assess the costs and benefits. Typically, research methodologies such as this either result in a cost or time savings, or open up new areas of value for the organization. In financial terms, profitability is improved either by increasing revenue or decreasing expense. This research is more geared to opening up new areas of value, which could lead to new sources of revenue.

Organizations have to invest a certain degree of effort in interacting with this research methodology before benefiting, as seen by the time taken by the experts to product the initial CLIOS representation and then review the results. Organizations will weigh the cost of doing this against the benefit. In doing this, a reasonable question is whether there is any cost savings – time and effort saved by using this research. There could be a large number of people working on market analysis and organizations may ask if that number could be reduced by using this research. This question was not explored in an experimental form in the three case studies, beyond interview feedback. Based on interview feedback, this research is likely not going to explicitly replace any existing processes. Although ad hoc review of blogs and other publications occurs, there is no single step in the typical process used by the experts which is replaced by this research. Instead, the additional value derived from this research is that it could open up new opportunities in the longer term and this value is what should be assessed when determining whether to apply this research in practice.

The value of this research is a combination of the value of the information being studied in the blogs and the people being used to interpret the results. Blogs contain a certain degree of information, and obviously if blogs in a certain domain are not generally trustworthy, then the research would not be valuable. However, experts are also required to interpret the results and an organization which uses this research would not benefit if the experts did not properly interpret the results. There is also a cost related to taking the wrong decision, and this research can help advise experts in making the right technology strategy decisions.

The overall cost for applying this research could be calculated by combining the time and effort required to execute the process and the degree of expert consulting required in interpreting the results and recommending actions. With further study, the degree of expert consulting can be reduced, which can improve the cost-benefit ratio of this process. The future study section in Chapter 9 covers some options for further automating this research.

8.2 Hypotheses and Research Questions Review

At the outset of this research, a set of research questions and hypotheses were developed and proposed to structure the research and assess whether the research meets the original research goals. This section provides a brief summary of how each hypothesis was studied in the course of this research. The hypotheses are each restated and then discussed, followed by a discussion of the research questions.

8.2.1 Hypotheses

H1: Blogs posts contain information which is both of interest and useful to subject matter experts in various domains.

This hypothesis was studied by asking each interviewee whether they currently consulted blogs, and also by reviewing the quantitative results from the interviews. Some experts did indicate that they already consulted blogs, but none did so in a formal or regular manner. The quantitative results categorized the findings from the blog analysis. The percentage of results which fell into the first two categories – known insights and worthy of investigation – showed that a sufficient volume of interesting and useful information was indeed contained in the blogs.

H2: The content of actual blog posts over time can be collected and summarized by using a statistical algorithm. Results with actual blog posts will be less sensitive to variation in the algorithm's parameters than results with random blog posts.

This hypothesis was addressed by the algorithms which were discussed in Chapter 3 and the associated software implementation which was used in this research. The algorithms basically collect information by doing periodic (ex. monthly) searches for keywords and summarize this information by proposing candidate links between keywords which are the output of latent semantic analysis and Google similarity distance calculations. For each month, a set of groupings were selected, and then those groupings of terms which occurred consistently over a number of months were selected to produce the candidate links. The sensitivity of these results was tested in the tests described in Chapter 4, where the algorithms were tested by adjusting parameters such as the number of months a link needs to occur to be selected, and also by comparing randomly constructed data against real data.

H3: Candidate changes to a system representation can be produced in a repeatable form

This hypothesis essentially proposes that a deterministic computer program can be written so that the same inputs to the program will produce the same candidate changes. It was tested by using the algorithm developed in testing H2, and applying explicit thresholds to the output so that the same candidate changes would result from the same data. The thresholds ensure that when a given set of data is analyzed, the results are always translated to the same set of candidate changes. For example, when a threshold of 50 months is chosen for the number of months which an anchor pair must appear, then every time the algorithm is run against the same data, it will produce the same set of candidate changes. If the threshold is altered, then the set of candidate changes may be altered.

H4: A reasonable number of accepted changes to the system representation will create insights into the domain which were not previously considered by technology strategy decision makers. These insights will inspire a variety of different technology strategy actions.

This hypothesis was tested by the interviews of experts in each case study. First, quantitative interviews categorized the candidate changes and confirmed that a reasonable number were indeed new and worthy of investigation. Next, the qualitative interviews explored the various technology strategy actions which the experts felt were inspired by the results. Examples of actions included writing proposals for research, entering new markets with products, and making technology policy decisions. These results were documented in each of the case study chapters as well as in this chapter. This research did not study non-technology related domains and thus has not tested whether this research can only benefit technology related domains. The interview feedback did not provide any specific evidence that this research should be limited to technology domains.

H5: The methodology developed in this study is independent of domain choice.

The use of three domains in this research showed that this methodology could be applied to a variety of domains. We did not test whether there are any domains for which this research could not be applied, but no limits to which domains could be studied were found in this research. Obviously, further study can be done to both explore more domains as well as to set limits on the type and scope of acceptable domains for this research.

8.2.2 Research Questions

The hypotheses served to structure the set of work required to answer the two research questions posed in Chapter 1. These questions are restated here with a brief discussion on how they have been answered in this research.

The two research questions are:

1. Can one develop a method of scanning social information systems that better informs people making technology strategy decisions?

2. Can improved information retrieval and representation of social information systems lead to different outcomes for technology strategy decision makers?

The first research question involves two main aspects – developing a methodology and then showing that it results in "better" information for a class of individuals. The methodology for scanning social information systems was indeed developed in this research, as discussed in the hypotheses review. The best evidence that was collected to determine whether the method "better informed" is the interview feedback. In the interview feedback, there is definitely evidence that the information produced was useful to the experts making technology strategy decisions. However, there were no experiments done which would demonstrate that this information is better than information available through other sources. That direct comparison is saved for future study.

The second research question involves the outcomes – which actions and which results are produced by this research. This question was answered by the interview feedback summarized in the first section of this chapter. The various actions were discussed and the potential results of these actions were described by the experts in each case study. The fact that a reasonable percentage of the insights found in this research were both unknown and had face validity meant that this research could lead to different actions and results than those originally planned by the experts.

8.3 Chapter Summary – Overall Results

This chapter described how this research could be used in practice by summarizing the results from the interviews conducted in all three case studies. This chapter also confirmed that the research had addressed each of the hypotheses identified at the outset of the dissertation. The results showed that this research can be applied in a variety of forms. It can be used to support intuition with fact and perception based analytics. The insights can be used to bridge understanding across roles and regions, and then map patterns across domains or subdomains. In certain cases, expected links did not occur and this requires more refined study to identify those cases where missing links indicate a new insight which the expert is not aware of. Understanding past events through the time series analysis was found useful by the stakeholders. The time series combined with the system-based analysis used, experts were informed on how to execute upon the ideas and insights they gained. Based on all of these different types of value gained from the research, the application of this research should be driven by a value-based assessment rather than a cost-benefit analysis which is based purely on savings against current processes and effort. Finally, the combination of the algorithms, software implementation, and interviews, served to test each of the 5 hypotheses identified in Chapter 1 and answer the research questions proposed at the outset of this research.

Chapter 9 Conclusion

In concluding this dissertation, the specific intellectual contributions are outlined, and lead directly into the overall impact of this research. The overall impact of this research provides a link back to the original motivation for this research and places this research in a broader context. This leads to a set of opportunities for future study which build on those contributions.

9.1 Contribution

This research provides an intellectual contribution on four levels. First, there is progress made in the underlying theory of social information systems, and this is further split into two distinct areas – statistical analysis algorithms and system-based analysis such as the CLIOS process. Next, there is a set of contributions specifically around the study of blogs, and these are also potentially applicable to other information sources. The actual implementation of the theoretical contributions of this research serves as a contribution in itself, as it provides an implementation platform for studying other domains. Finally, this research was conducted as part of a broader program on technology forecasting at MIT and the Masdar Institute, and this research has contributed to progress in the field of technology forecasting.

9.1.1 Theoretical contribution to the study of "social information systems"

In Chapter 1, the notion of "social information systems" was defined as systems which allow participants to produce content, share it with others, and allow readers of the content to provide their own commentary or integrate others' content into their own content. Blogs were chosen as a representative social information system. The main theoretical contribution of this research is in demonstrating that information does not have to be read in its entirety to be used. This is a fundamental contribution because the nature of social information systems is that content is easy to produce, and thus high volumes of content are often produced around any topic of interest. These high volumes of content risk not being utilized if individuals are required to read the entire set of information. Having a means to analyze information and apply new knowledge to help make decisions without actually reading the entire set information allows this vast array of information to be utilized.

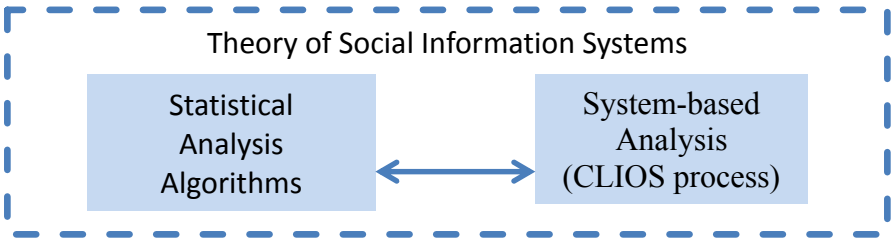

Figure 9.1. Theoretical contributions to social information systems

In Figure 9.1, the two subparts of the theoretical contribution are depicted. The methodology developed in this research made a contribution to both the area of statistical analysis algorithms as well as system-based analysis, specifically with contributions related to the CLIOS process. Figure 9.1 depicts these two areas as connected by a two-way arrow because another major

contribution of this research is the mapping between the statistical algorithms and the system-based analysis.

Specific algorithms for statistical analysis

The algorithms developed to conduct the bibliometric blog analysis were themselves a contribution of this research. They provided a novel method for using search to observe the co-occurrence of terms over time and then derive insights. The selection of LSA and Google similarity distance could be replaced with other methods, but the overall approach of selecting specific terms from a system representation, splitting results into periodic (ex. monthly) sections, grouping related terms, and identifying trends over time, can be used regardless of specific implementation decisions.

Enhancement to system-based analysis such as the CLIOS process

Mapping the statistical approach to the CLIOS process allowed for specific statistical insights to be used to drive specific insights about the system. For example, the use of keywords to indicate performance related topics such as "cost" allowed performance measures to be identified and the use of LSA-based concepts allowed links between components to be identified.

Independent of the mapping of statistical analysis to the CLIOS process, there are two additional areas of contribution to the field:

1 – A new use of the system representation phase of the CLIOS process, in the context of projects in which the entire CLIOS process was not being conducted and very specific goals around system representation were being pursued.

2 – A new feature of the CLIOS process which allows for re-representation based on an initial representation of the system and the incorporation of new insights which suggest changes to the representation. This new feature requires both the ability to collect new insights on a regular basis – achieved via the statistical bibliometric blog analysis method – as well as the methodology for accepting changes into the system representation – achieved via the interview process. Both of these aspects were described in detail in Chapter 3.

9.1.2 Information sources – studying blogs, and applying more broadly

Another area of contribution of this research is in the study of blogs, as well as the opportunity to take what has been learned by studying blogs and apply it more broadly to other information sources. In terms of studying blogs, Chapter 2 covered the current state of the art and exposed a lack of sophistication around current methods for studying blogs. This research contributes to the study of blogs by providing a methodology for doing periodic searches on terms contained within blog posts, storing those results, and analyzing them for patterns and trends. This contribution can be used in other research endevors where the researcher seeks to study blogs as an information source.

The method used to study blogs can be applied more broadly to other information sources. Figure 9.2 depicts other possible information sources, both less structured and more structured

than blogs. Less structured sources include emails and microblogs. E-mails were suggested by experts in the Kenya study as another form of communication used to discuss broadband expansion. Microblogs, such as Twitter, are emerging as another communication vehicle for certain types of conversation. More structured sources include journals and patents. Many of the same principles around periodic searches for terms and analysis by grouping related terms can be applied to these sources as well. Effectively, any information source which can be searched over time, and which contains some significant volume of contribution, can be studied using the methods developed in this research.

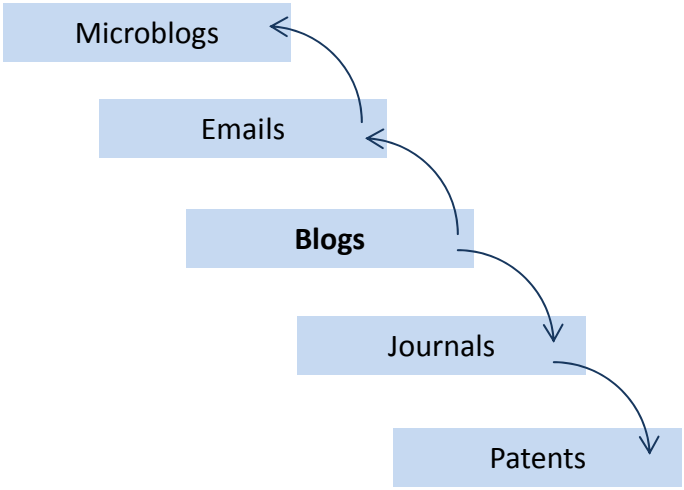

Figure 9.2. Blogs as one of many possible information sources

9.1.3 "Cookbook" implementation for other domains

The set of steps used to apply the statistical algorithms in a software implementation and then conduct structured interviews to review the results in both a quantitative and qualitative fashion is also in itself a contribution because it provides a pattern for implementing this research in other domains. This contribution can be thought of as a process contribution similar to the notion of conducting clinical trials on pharmaceuticals. In the case of the pharmaceutical industry, the process of systematically selecting a population, administering the appropriate medications, comparing against a population who has taken placebo medication, and then reviewing results with the relevant regulatory body is what made it possible to deliver pharmaceuticals to market. In this research, the set of process steps developed provides a pattern that allows information to be analyzed and used in practice. Although this was scoped to technology strategy in this research, it certainly has the potential to be used in non-technical arenas as well. Interviewees mentioned the possibility to use this research for studying political issues, conducting sentiment analysis on particular topics, and any other areas which require analysis of a widespread amount of public discussion.

9.1.4 Technology Forecasting

This study was conducted as part of broader research on technology forecasting in a joint endeavor between MIT and the Masdar Institute of Science and Technology. As such, the contribution of this research towards the field of technology forecasting is that it provides the ability to track the interrelationship of terms over time, and it provides the ability to track the changes in a system context over time. Both of these contributions allow prediction of future behavior with respect to the various interrelated factors inside a specific domain. Although prediction on its own is not a significant contribution, having high quality predictions which experts trust is a significant contribution. There are many approaches to technology forecasting, and this research was one project within the overall technology forecasting area. Other methods of forecasting technology which were studied are taxonomy building which constructs a hierarchical set of topics within a domain, and term discovery which identifies new terms to be studied within a domain.

9.2 Impact

The contributions discussed earlier as well as the doors which have been opened for future study, suggest five broad areas in which this research can have an impact beyond the specific case studies. This research paves the way for harnessing information without actually reading the information, which then provides earlier awareness of information required to predict future direction and also connects experts and amateurs in the same domain who may not have been connected before. Another major impact this research can have is the ability to collect public perception and organize a view of a domain based on public perception rather than pure fact. Finally, the methodology developed in this research is a tool which itself can have an impact on researchers who are conducting similar studies.

Before discussing each specific area of impact, it is important to review the original motivation for this research. Volume of information is growing rapidly, and just as machines were critical to the industrial economy, information processing is the key growth factor for the knowledge economy. Social information systems are a class of information which requires particularly novel methods for analysis, and blogs are a representative source of such information. Complex decisions which involve multiple parts of a system will benefit from having the ability to gather and review information from these emerging sources. In this context, the following areas are where this research can have a particularly powerful impact.

9.2.1 Ability to harness information without reading information

The ability to harness information without actually reading the information means that individuals can focus their effort on reviewing critical information and making decisions rather than spending time reading every piece of information that is produced. This requires trusting automated "agents" to review and analyze the information, and this research provides an implementation of such an agent. This is a fundamental shift from prior milestones in the growth of the Internet. The Internet was first focused on connecting information sources, and then on making it easier for anyone to publish information to the Internet. Now, it is critical to be able to rapidly review all of the produced information, and drive action based on the insights which are found. This is similar to the evolution of stock investment systems – initially, investors relied on

reviewing all information and making explicit decisions, but in recent time, it has become possible to set up investment rules which automatically scan financial data and act on behalf of the investor. Other options for collecting information and intelligence from a broad group of Internet users are often based on serendipitous encounters or on single question polls, and the system-based analysis conducted in this research allows for a broader set of insight to be derived and a more systematic set of actions to be taken.

9.2.2 Earlier awareness of information to predict future direction

Being able to harness information automatically allows earlier awareness of information, and can also lead to predictions of future direction. The introduction established that blogs make information easy to publish, and the case studies demonstrated certain situations where insights were seen earlier in blogs than in other forums. The automatic analysis that can be done for an expert can produce relevant insights that are captured earlier than if the expert had the time to read everything on the Internet. This is because not only are the automated tools reading and summarizing the information, they are also connecting concepts which may or may not be explicitly discussed in the blogs. For example, links between different components, or the link between a specific component and an indicator term such as cost, are often found through multiple steps of analysis. This could also lead to the introduction of analytics into the web browsing experience. Social information systems allow the reader to automatically jump from one article to another based not just on content but also on people – for example, if an individual comments on a blog entry, it's possible to immediately go to their own blog. This research promotes the possibility that a reader may also be able to go from a blog on one topic to a view of the overall set of related concepts, without relying on the blog author to create the links themselves.

9.2.3 Connect experts and amateurs

In any domain, there is a spectrum of individuals at different levels of expertise, but there are often very different forums of discussion for the seasoned experts and the casual observers or amateurs. For example, conferences and journals are the forums used by experts in renewable energy, however there are many news articles and online articles which include viewpoints from those with a casual interest. This is especially true in this example because renewable energy is a politically relevant topic in many countries. The viewpoints of experts and amateurs are not often connected in a detailed manner, and this research allows the possibility to link the two and share information in both directions. Amateurs have the ability to gain deeper insights into the domain without having to process all of the information published by experts. Experts have the ability to gather the aggregate opinion of the interested community, and potentially predict which seemingly amateur ideas will be the expert ideas of the future.

9.2.4 Build a "Perception Taxonomy" of a domain

In complex systems, the line between reality and perception is not always clear, and the two often influence each other. Public perception has the ability to influence reality, and there is often an intentional desire to affect public perception. Blogs are well positioned to bridge between fact and perception and this research provides a means for understanding public

perception and the interaction of physical components of a complex system with perception. In market-based systems, consumers make their selections based on perception and so understanding what drives perception is critical.

The notion of taxonomy has been used in technology forecasting to help categorize the various elements of a domain into a hierarchical form. Taxonomies identify which terms are specific instances of a more general term, and which terms are higher level categories for more specific terms. This results in a tree-like depiction of a domain where the most general topics are at the top, and greater degrees of specificity result as the branches of the tree are traversed. These taxonomies are based on fact – they are meant to depict the physical and technical reality underlying the domain. The classification systems used for animals or for medical fields are good examples of taxonomies.

This research allows us to define a new concept, a "perception taxonomy." A perception taxonomy is one which categorizes the topics within a domain based on the public perception of their importance or prominence. Perhaps the most telling example of a categorization which would be different in a perception taxonomy is the discussion in the renewable energy case that ethanol was a more popular and general term than biofuels. The technical reality is that biofuels are a general category under which ethanol is one type of biofuel. However, the public perception as indicted by blog discussion is that ethanol is a high level category. Public perception could be as critical as reality in certain domains, and the ability to generate this "perception taxonomy" opens up the ability to understand needs, interests or simply misinformation that is prevalent in the public.

9.2.5 Extensions to research methodologies in practice today

This research provides a set of extensions and new uses to the CLIOS process which were discussed in section 9.1. The CLIOS process is a useful tool for researchers to conduct in depth analysis of a domain and inform strategic decisions. The extensions made in this research make an impact on both users of the CLIOS process as well as other researchers who are looking for new ways to study domains that bridge quantitative and qualitative analysis methodologies. On a more general level, this research provides a roadmap for researchers to incorporate text analysis tools and packages with research methods they are already employing. This research also demonstrates that it is not a requirement of statistical algorithms to produce definitive results, and it can be sufficient to produce a set of candidate findings from which an interview process can select the appropriate final results.

9.3 Future study

Specific areas for future study were referenced throughout Chapters 8 and 9, and this section provides a comprehensive list of opportunities for future study. The areas are summarized in Table 9.1 and then discussed in more detail following the table.

Area	Details
Broader technical	Choice of LSA and Google similarity distance as the

experiments	analytical methods used to analyze blogs can be revisited More/less frequent data collection "Term discovery" – identifying new terms to search Condense set of candidate changes to make interviews efficient
Different approaches to including system view	Apply method to policy levers, other aspects of system representation Consider other phases of CLIOS process beyond representation – design and selection of alternatives Consider removing links or aspects from the system representaiton Consider other methods (system dynamics, …) Consider no system view, pure stats, or generated system view
Interview process enhancements	Schedule more time with experts More automated results and survey style technique
Broader data sources and quality of sources	Other sources of blogs beyond Google Apply research to journals, patents, news, microblogs, other sources Merge results from multiple sources
Exhaustive study of multiple domains and roles to assess differences	Define parameters to compare domains based on topic area and scope Use survey technique referred to above to quickly gather feedback from many domains Assess interview feedback for various roles and assess differences
Studying blogs inside an organization or community	Use data from corporate social networks Focus on blogs from a particular set of validated authors

	Include authors as a parameter to 'score' blogs
Recommending and Assessing Actions of Experts	Identify explicit decisions to be made, and develop algorithm to suggest explicit actions based on results Use taxonomy to help filter results to the key actions to be taken Assess the results of actions taken Assume experts involved in CLIOS initial representation, could reconsider this
Prediction	Explicit prediction of future behavior Assessment over time of whether predictions came true by running experiments now and then same experiments a year later

Table 9.1. Areas for future study

9.3.1 Broader technical experiments

The first and most basic area of future study is to simply continue the progress made in this research in the specific technical experiments described in Chapter 4. Two analytical methods were chosen – latent semantic analysis and Google similarity distance – but others exist which can be used to process the results of the blog searches. The methods chosen in this research did not account for the number of occurrences of a keyword in a blog post, and this is one example of an improvement which may be offered by choosing another method. Inserting another analytical method would not change the other steps in the methodology; it would simply change the method by which the raw hit counts were transformed to candidate changes. This research also chose to use a monthly frequency of data collection, and this frequency can be adjusted to be more or less frequent. One possibility would be to do daily or hourly searches and examine the potential for a "real-time" dashboard style of presenting results to experts.

This research only used those terms which were predefined by the experts as part of the initial CLIOS representation. There are various methods for term discovery which could be used to produce additional terms for study related to the terms found in the initial CLIOS representation. Also, each component in the system representation was used as one term – a more comprehensive approach would be to select a set of terms which together represent a given component in the system representation, and build an algorithm for aggregating results for all terms that relate to a component into a view of that component's result.

The candidate changes presented to experts were a complete list of changes that the algorithm produced. This list could definitely be condensed if further algorithms were developed.

Condensing and prioritizing this list would lead to more efficient interviews as well as more automated changes to the system representation.

9.3.2 Different approaches to including system view

The basic mapping of certain analytical methods to certain aspects of the CLIOS system representation process leaves opportunities for further work on the following dimensions:

- Additional CLIOS representation aspects can be included as candidate changes – for example, identifying a component as a policy levers or institutional actors.

- Additional steps in the CLIOS process can be informed by the insights found in the blog analysis – for example, the design and selection of strategic alternatives that an expert can choose between

- Removing links instead of adding them to the representation

- Other system-based analysis methods beyond the CLIOS process – network analysis and system dynamics are two other examples

- Instead of generating an initial representation via expert interviews, the process can start by automatically generating the representation

- Pure statistical insights, without a system-based analysis, can be examined to determine the exact effect of including a system-based analysis

9.3.3 Interview process enhancements

The interview process evolved as each case study progressed, and further case studies can further enhance the interview process. In terms of interview time and structure, two directions can be pursued. Scheduling more time with experts may allow for more in depth review of the candidate changes and more examination of the possible actions that are taken based on the insights. This includes potential follow-up interviews to assess actions after some period of time. The other extreme would be to further automate the selection of candidate changes, and use less time with experts, relying instead on a survey style technique which quickly gathers feedback on the results from a broad population.

9.3.4 Broader data sources and quality of sources

As discussed in the contribution section, the various information sources beyond blogs all remain possibilities to study beyond this research – microblogs, emails, journals, patents and other sources. Each information source has a unique set of characteristics, but the generic search algorithms and implementation developed in this research can potentially be re-used across broader sources. Any source which can be searched via the web can be included with little effort into the implementation produced in this research. The level of structure and the amount of meta-data available changes with each information source, and studying a broader set of sources could lead to better utilizing the meta-data associated with an information source. Even with

blogs, other sources beyond Google's blog search engine can be used, to examine the impact of selecting Google as the source for blog searches. There is also the possibility to merge results from multiple sources, to examine the aggregate insights that can be acquired.

9.3.5 Exhaustive study of multiple domains and roles to assess differences

Only three domains were studied in this research. They were compared informally in Chapter 8, but more rigorous comparisons are possible. Defining parameters to compare domains based on topic area and scope would allow for more domains to be selected, and the applicability of this research in different domains could be studied. There may be a certain scope and topic area for which this research has especially profound impact. The survey technique discussed in the interview process subsection could be used to quickly gather feedback from many domains. Similarly, a more exhaustive study of the various roles which experts hold could also be done, to examine the applicability of this study for different roles.

9.3.6 Studying blogs inside an organization or community

This research focused on publicly available blogs, but as the cloud computing study revealed, corporate social networks are becoming increasingly popular. These blogs have many of the same technical characteristics and thus the same implementation can be used. There are also additional characteristics of corporate blogs which can be leveraged. Corporate blogs are only written by employees of the company, and many companies have a corporate directory which provides well-structured information about every individual in the company. Thus, further algorithms can be developed to influence the analysis based not just on the content of the blogs but on the expertise and role of the blog author.

For public blogs, there is the option to focus on blogs from a particular set of validated authors or in a particular community or region. Focusing on a certain community of authors could also be done by using blog author characteristics to 'score' blogs based on relevance. This would give more prominence to those blogs which had relevant authors, but would not rule out those blogs whose authors were not known as expects.

9.3.7 Recommending and Assessing Actions of Experts

Rather than simply providing insights to experts and letting the experts determine their own actions, there exists the possibility to recommend actions. This can be done by identifying different action types which an expert can pursue, and then developing algorithms which map statistical results to these action types in the same manner that statistical results were mapped to the different aspects of the system representation. Taxonomy development can help filter results to focus on the key actions to be taken, by exposing the set of high level issues involved in the domain.

In terms of assessing the actions of experts, this research focused on a one-time interview with the experts, but further study could conduct multiple interviews which would allow for assessment of actions. For example, if an expert reviews the results and takes a specific action with an assumption that a particular topic area will grow, that assumption can be assessed by

conducting the analysis and interview again at a later date. This leads directly into the area of prediction.

9.3.8 Prediction

Perhaps the least defined and furthest reaching area of future study is in the explicit prediction of future behavior. The trends found by the statistical algorithms in this research can be combined with known methods for curve-fitting, to produce assertions about future behavior of different topics and the links between those topics. This also allows for an assessment over time of whether predictions came true by running experiments now and then same experiments a year later, and also interviewing experts later on to assess whether predictions held true.

9.4 Chapter Summary – Conclusion

This chapter concluded the dissertation by presenting a set of intellectual contributions, then discussing future study which could build on those contributions, and ends with areas of broader impact which are more forward thinking than the explicit contributions of this research. The intellectual contributions include enhancements to the theory of social information system around statistical analysis methods and extensions to the CLIOS process, enhancements to the study of blogs and the ability to study broader information sources, a process for implementation which can be used as a pattern for studying other domains, and contribution to the various methods for technology forecasting. Opportunities for future study exist across specific extensions to each phase of the research, broadening the sources and domains studied, and getting into new use cases such as recommending actions and prediction. This research set the stage for five areas of broader impact: automatically reviewing information without reading it, gaining early awareness of insights, connecting experts and amateurs in a field, building a "perception taxonomy" of a field, and building research tools for broader research.

Appendix 1 Cloud Computing Term Clusters

Cloud Computing – Biz Models for Producers subsystem

Installation-subscription (12,27) -
Frequent addition:
cost (6), comes in and out
Occasional additions to the concept:
time to market (28)
open cloud (17)
hardware as a service (9)
land and sky (14)

Perpetual License - Cannibalization (20,4)
Disappears sometimes, gone after mid-2008
Frequent:
Walled Garden (32)
Occasional:
Utility, Utility Computing (31,30)
Up Front Commitment (29)

Walled Garden - Cannibalization (32,4)
Gone after mid-2008, back in late 2009
Frequent:
Perpetual License (20)
Occasional:
Utility Computing, Land and Sky (30,31,14)
Up Front Commitment (29)
Pilot Usage (21)

Agility - Database as a Service (1,7)
only showed up once in 2005 (noise?)
Frequent: none
Occasional: none

Open Cloud - Social Networks (17,24)
Starts in late 2006, picks up in late 2007
Frequent:
Time to Market (28)
Cannibalization, Software as a Service (3,25) - last 4 months
Occasional:
Infrastructure as a Service, Time to Market, On Demand (11,28,16)
Always On (2)
Utility Computing, Utility, On Demand (31,30,16)

On Demand - Social Networks (16,24)
Gone for most of the time, or is inside a large concept
Frequent:
Time to Market (28) - last 3 months
Occasional:
Utility, Utility Computing (30,31)
Storage as a Service, Pay Per Use (26,19)
Internal Cloud, Managed Services (13,15)
Perpetual License (20)
Hardware as a service, Storage as a service (9,26)
Platform as a service, Shared Infrastructure (22,23)
Time to Market, Platform as a Service (28,22)
Infrastructure as a service, Software as a service (11,25)

Utility - Utility Computing (30,31)
Frequent:
Always On (2)
Time to Market, Up Front Commitment (28,29) - shows up for last 7 mos
Occasional:
Social Networks, On Demand (24,16)
Perpetual License, Cannibalization (20,4)
Shared Infrastructure, Database as a Service, Managed Services (23,7,15)
Database as a service, Ecosystem, Agility (7,8,1)
Database as a service, Internal Cloud, Always On (7,13,2)
Platform as a Service (22)
Internal Cloud (13)
Cannibalization, Walled Garden (4,32)
Time to Market, Social Networks (28,24)

Cloud Computing – Audiences – Industry, Country subsystem

Consistently appear linked 2005-2009
enterprise and data+center (14, 8)
India and broadband (18, 3)
communications and China (5, 4)
data+center and China (8, 4)
developers and content+anywhere (10, 6)
device+access and content+anywhere (12, 6)
enterprise and China (14, 4)
social+networks and enterprise (26, 14)
mobile and enterprise and communications (20, 14, 5)
social+networks and mobile and communications (26, 20, 5)

Concepts
developing+nations and developed+nations (11, 9)
2006 to 2008, Dec 2009
Bandwidth – handful
Regulation – twice
Russia, protocol – once
Mid market – handful 2
Shared infrastructure – once
Content anywhere – once
Healthcare industry – once in dec 2009

regulation and broadband (22, 3)
Brazil – frequent
SMBs – twice with Brazil in 2006
Shared infrastructure, Content anywhere – once
India, mid market, financial industry – once
Customer relationship management, Brazil, mid market – once
Developed nations, bandwidth – once
Startups, customer relationship management, healthcare industry – once
Mid market, Brazil, financial industry – once
India, Brazil – twice
Encryption – twice

Social networks and mobile and enterprise and communications (26, 20, 14, 5)
Financial industry – twice in 2006
Data center – once
China – twice in 2006/2007
India – once in 2006
Customer relationship management – handful in 2006/2007
Startups – once in 2008

shared+infrastructure and protocol (24, 21)
Broadband , developers – once in 2005
Encryption – once in 2006
Bandwidth – once in 2007
Developing nations – frequent 2007-2009
SMBs, Russia – freuqent 2009

device+access and content+anywhere (12, 6)
Communications – once in 2005 and 2009
Healthcare industry – once in 2006
Mobile – once in 2009

SOX and HIPPA (27, 17)
SMBs – handful, late 2005 to 2007
Brazil – three, late 2005, 2006
Healthcare industry – once
Russia – once
Encryption – once
Shared infrastructure – once
Developing nations – twice

shared+infrastructure and developing+nations (24, 11)
2006 onwards

Protocol – appears frequently
Broadband – handful
Bandwidth – handful
Encryption – once
SOX, HIPPA, SMBs – once
SMBs – handful
Russia – twice in 2009

protocol and developing+nations (21, 11)
2005-2007, 2009
Shared infrastructure – frequently
Regulation – once
Startups – once
Russia, mid market, content anywhere – once
Bandwidth, mid market – once
Russia – three times

developed+nations and bandwidth (9, 1)
2007 to 2008
Protocol – twice
Financial industry – once
Shared infrastructure – once
Financial industry – once
Mid market – twice
SMBs, Russia – once
SMBs, regulation – once
Encryption – three times in 2008
Regulation – three times in 2008
Broadband – twice

Cloud Computing – IT Delivery Channels subsystem

regulation and compliance (19, 4)
'reliability', 'distributed+databases‘ – once
Operating system – handful
Automation – handful
On demand, integration – once
APIs, Mashups – handful
Integraiton – once
'shared+infrastructure', 'operating+system', 'disk+drives‘ – once
Scalability, automation – once
'disk+drives', 'automation', 'scalability‘ – once
Virtualizaiton – once

security and device+access (22, 6)
Virtualization, web based computing – handful
Open standards – once
Mobility – once

web-based+computing and security and mobility (26, 22, 13)
Virtualization, device access – handful
Mashups, content anywhere – handful
Open standards, device access – once
Compliance – twice
Operating system – frequent after mid 2007
Scalability, open standards – once

mashups and automation (12, 2)
open+standards and integration (16, 10)
on-demand and integration (15, 10)
regulation and automation (19, 2)
mashups and compliance (12, 4)
Virtualization and open+standards (25, 16)
operating+system and integration (17, 10)

operating+system and open+standards (17, 16)
device access, integration, content anywhere – once
security, web based computing -

open+standards and on-demand (16, 15)
device access, content anywhere – couple times early on
virtualization, integration, APIs – handful throughout the timeframe
operating system – handful
mashup - once

Virtualization and device+access (25, 6)
security, web based computing – twice in 2006

Cloud Computing – Enterprise Value Capture subsystem

These subsystems are shown in a different format because they were created after the software tool was automated to calculate the anchor pairs and associated terms. The first set of terms are anchor pairs which appear together in larger clusters, and the next set of terms are anchor pairs which appear together in smaller clusters. The anchor pairs for smaller clusters are shown in bold and the associated terms are shown underneath each anchor pair, along with which months those associated terms appeared in the results.

open+standards and customer+relationship+management (14, 4)
security and device+access (22, 6)
open+standards and business+processes (14, 1)
security and cost (22, 3)
cost and business+web (3, 2)
dynamic+allocation and customization (7, 5)
social+networks and cost (23, 3)
hosting and customer+relationship+management (9, 4)
customer+relationship+management and business+processes (4, 1)
dynamic+allocation and customer+relationship+management (7, 4)
outsourcing and managed+services (17, 12)
social+networks and privacy (23, 18)
privacy and business+web (18, 2)
security and privacy (22, 18)
security and business+web (22, 2)
time+to+market and business+processes (24, 1)
time+to+market and customer+relationship+management (24, 4)
time+to+market and open+standards (24, 14)
privacy and cost (18, 3)
hosting and customization (9, 5)

security and device+access (22, 6) - {
[‘scalability', 'global+marketplace', 'managed+services']": ['Mar09'],
[‘dynamic+allocation', 'business+web', 'cost']": ['May06'],
[‘cost', 'social+networks']": ['Mar07'],
[‘privacy', 'customer+relationship+management']": ['Aug06'],
[‘privacy', 'business+processes']": ['Apr06'],
[‘organiational+change', 'scalability', 'dynamic+allocation']": ['Dec09'],
[‘dynamic+allocation', 'organiational+change', 'global+marketplace']": ['Sep09'],
[‘hosting', 'customization']": ['Feb07'],
[‘hosting', 'privacy']": ['Nov07'],
[‘business+processes', 'privacy']": ['Apr06'],
[‘business+web']": ['Jun05'],
[‘global+marketplace', 'scalability']": ['Aug08'],

[‘scalability']": ['Sep08', 'Nov08', 'Dec08'],
[‘cost', 'privacy', 'social+networks']": ['Nov06']}

security and cost (22, 3) - {
[‘device+access', 'dynamic+allocation', 'business+web']": ['May06'],
[‘privacy', 'social+networks', 'device+access']": ['Nov06'],
[‘device+access', 'social+networks']": ['Mar07'],
[‘business+processes', 'privacy', 'business+web']": ['Dec05'],
[‘business+web', 'privacy', 'social+networks']": ['Nov07', 'Jan08', 'Jan09', 'Feb09', 'Aug09'],
[‘privacy']": ['Jun06'],
[‘business+web']": ['Oct05'],
[‘business+web', 'privacy']": ['Sep05', 'Jun06', 'Sep06', 'Feb07', 'Jun07', 'Jul07', 'Aug07', 'Sep07', 'Nov08'],
[‘privacy', 'business+web']": ['Sep05', 'Jan06', 'Mar07', 'May07']}

cost and business+web (3, 2) - {
[‘security', 'privacy']": ['Sep05', 'Jun06', 'Jul07'],
[‘device+access']": ['Nov05'],
[‘customization', 'dynamic+allocation']": ['Mar07'],
[‘security', 'privacy', 'social+networks']": ['Aug09'],
[‘privacy', 'business+processes']": ['Dec05'],
[‘device+access', 'business+processes', 'time+to+market']": ['Jun05'],
[‘privacy', 'security', 'social+networks']": ['Nov07', 'Jan08', 'Jan09', 'Feb09'],
[‘managed+services', 'outsourcing', 'customer+relationship+management']": ['Aug08'],
[‘privacy', 'security']": ['Sep05', 'Jan06', 'Sep06', 'Feb07', 'Mar07', 'May07', 'Jun07', 'Aug07', 'Sep07', 'Nov08'],
[‘security']": ['Oct05'],
[‘privacy']": ['Oct05', 'Jan07', 'Apr07'],
[‘social+networks', 'privacy', 'customization']": ['Feb08'],
[‘device+access', 'privacy', 'social+networks']": ['Nov06'],
[‘device+access', 'security', 'dynamic+allocation']": ['May06'],
[‘customization', 'privacy']": ['Dec07'],
[‘open+standards', 'dynamic+allocation']": ['Jun05'],
[‘security', 'business+processes', 'privacy']": ['Dec05'],
[‘privacy', 'social+networks']": ['Jun08'],
[‘customization', 'open+standards', 'social+networks']": ['Jun08']}

outsourcing and managed+services (17, 12) - {
[‘productivity', 'hosting']": ['Jul09'],
[‘customer+relationship+management', 'regulation', 'hosting']": ['Jun09'],
[‘cost']": ['Jan09'],
[‘productivity']": ['Jan06', 'Aug08', 'Sep08', 'Dec08', 'Jan09'],
[‘customization', 'scalability', 'customer+relationship+management']": ['Mar09'],
[‘cost', 'customer+relationship+management', 'business+web']": ['Aug08'],
[‘movement+of+applications']": ['Jul07'],

[‘privacy']": ['Jan07'],
[‘customization', 'regulation', 'movement+of+applications']": ['May06'],
[‘regulation']": ['Jun08'],
[‘productivity', 'movement+of+applications', 'scalability']": ['Mar07'],
[‘open+standards']": ['Sep05'],
[‘dynamic+allocation']": ['Jun09']}

social+networks and privacy (23, 18) - {
[‘open+standards', 'managed+services', 'business+web']": ['Apr07'],
[‘cost', 'customization', 'business+web']": ['Feb08'],
[‘business+web', 'cost']": ['Jun08'],
[‘cost', 'customization']": ['Mar08'],
[‘business+web', 'managed+services', 'productivity']": ['May07'],
[‘security', 'cost', 'device+access']": ['Nov06'],
[‘business+web', 'cost', 'device+access']": ['Nov06'],
[‘business+web']": ['Sep07'],
[‘outsourcing', 'customer+relationship+management', 'business+processes']": ['Nov05'],
[‘cost']": ['Mar09', 'Jul09'],
[‘dynamic+allocation', 'open+standards', 'business+web']": ['Oct06'],
[‘business+web', 'security', 'cost']": ['Nov07', 'Jan08', 'Jan09', 'Feb09', 'Aug09'],
[‘customization', 'cost']": ['Sep08']}

organiational+change and long-term+commitments (16, 11) - {
[‘OpenSocial']": ['Oct06'],
[‘global+marketplace', 'labor+efficiency']": ['Apr07', 'Jun07', 'Aug07'],
[‘labor+efficiency']": ['Dec06', 'May07', 'Oct07', 'Dec07'],
[‘global+marketplace', 'movement+of+applications']": ['Jul08'],
[‘movement+of+applications', 'labor+efficiency', 'global+marketplace']": ['Jun07'],
[‘regulation', 'scalability']": ['Dec06'],
[‘labor+efficiency', 'regulation']": ['Nov06'],
[‘global+marketplace']": ['Mar06', 'Apr06', 'Sep07']}

privacy and business+web (18, 2) - {
[‘productivity']": ['Jul07'],
[‘social+networks', 'managed+services', 'productivity']": ['May07'],
[‘dynamic+allocation', 'customer+relationship+management']": ['Aug06'],
[‘customization']": ['Nov07'],
[‘social+networks']": ['Sep07'],
[‘cost']": ['Oct05', 'Jan07', 'Apr07'],
[‘outsourcing', 'hosting']": ['Jan07'],
[‘dynamic+allocation', 'social+networks', 'open+standards']": ['Oct06'],
[‘security', 'social+networks', 'cost']": ['Aug09'],
[‘cost', 'device+access', 'social+networks']": ['Nov06'],
[‘security']": ['Jul06', 'Aug06', 'Oct08'],
[‘cost', 'business+processes']": ['Dec05'],

[‘social+networks', 'cost', 'customization']": ['Feb08'],
[‘cost', 'social+networks']": ['Jun08'],
[‘customization', 'cost']": ['Dec07'],
[‘security', 'cost', 'social+networks']": ['Nov07', 'Jan08', 'Jan09', 'Feb09'],
[‘security', 'cost', 'business+processes']": ['Dec05'],
[‘security', 'cost']": ['Sep05', 'Jun06', 'Sep06', 'Mar07', 'Jun07', 'Jul07', 'Sep07', 'Nov08'],
[‘open+standards', 'social+networks', 'managed+services']": ['Apr07'],
[‘cost', 'security']": ['Sep05', 'Jan06', 'Feb07', 'May07', 'Aug07']}

long-term+commitments and global+marketplace (11, 8) - {
[‘organiational+change', 'movement+of+applications']": ['Jul08'],
[‘labor+efficiency', 'OpenSocial']": ['Mar08'],
[‘labor+efficiency']": ['Oct07'],
[‘regulation', 'labor+efficiency']": ['Nov07'],
[‘movement+of+applications', 'labor+efficiency', 'organiational+change']": ['Jun07'],
[‘movement+of+applications']": ['Oct07', 'May08'],
[‘organiational+change']": ['Mar06', 'Apr06', 'Sep07'],
[‘regulation']": ['Aug06'],
[‘organiational+change', 'labor+efficiency']": ['Apr07', 'Jun07'],
[‘customization', 'hosting', 'outsourcing']": ['Feb06'],
[‘social+networks', 'open+standards']": ['Aug05'],
[‘labor+efficiency', 'organiational+change']": ['Aug07']}

security and privacy (22, 18) - {
[‘hosting', 'device+access']": ['Nov07'],
[‘business+web', 'cost']": ['Sep05', 'Jun06', 'Sep06', 'Feb07', 'Mar07', 'May07', 'Jun07', 'Jul07',
'Aug07', 'Sep07', 'Nov08'],
[‘cost', 'business+web']": ['Sep05', 'Jan06'],
[‘business+web', 'cost', 'social+networks']": ['Nov07', 'Jan08', 'Jan09', 'Feb09'],
[‘device+access', 'business+processes']": ['Apr06'],
[‘cost']": ['Jun06'],
[‘cost', 'business+processes', 'business+web']": ['Dec05'],
[‘cost', 'social+networks', 'device+access']": ['Nov06'],
[‘business+processes', 'device+access']": ['Apr06'],
[‘business+web']": ['Jul06', 'Aug06', 'Oct08'],
[‘business+web', 'social+networks', 'cost']": ['Aug09'],
[‘device+access', 'customer+relationship+management']": ['Aug06']}

security and business+web (22, 2) - {
[‘cost']": ['Oct05'],
[‘privacy', 'cost', 'social+networks']": ['Nov07', 'Jan08', 'Jan09', 'Feb09'],
[‘cost', 'privacy']": ['Jan06', 'Jun06'],
[‘device+access']": ['Jun05'],
[‘privacy']": ['Jul06', 'Aug06', 'Oct08'],
[‘cost', 'business+processes', 'privacy']": ['Dec05'],

[‘privacy', 'social+networks', 'cost']": ['Aug09'],
[‘privacy', 'cost']": ['Sep05', 'Sep05', 'Sep06', 'Feb07', 'Mar07', 'May07', 'Jun07', 'Jul07', 'Aug07', 'Sep07', 'Nov08'],
[‘device+access', 'dynamic+allocation', 'cost']": ['May06']}

privacy and cost (18, 3) - {
[‘customization', 'social+networks']": ['Mar08', 'Sep08'],
[‘security', 'business+processes', 'business+web']": ['Dec05'],
[‘security', 'business+web']": ['Jan06', 'Jun06'],
[‘business+web', 'device+access', 'social+networks']": ['Nov06'],
[‘social+networks', 'customization', 'business+web']": ['Feb08'],
[‘hosting']": ['Jul07'],
[‘security', 'social+networks', 'device+access']": ['Nov06'],
[‘business+web', 'business+processes']": ['Dec05'],
[‘security']": ['Jun06'],
[‘social+networks']": ['Mar09', 'Jul09'],
[‘business+web']": ['Oct05', 'Jan07', 'Apr07'],
[‘customization', 'business+web']": ['Dec07'],
[‘business+web', 'social+networks']": ['Jun08'],
[‘customization', 'hosting']": ['Apr09', 'Jul09'],
[‘business+web', 'security']": ['Sep05', 'Sep05', 'Sep06', 'Feb07', 'Mar07', 'May07', 'Jun07', 'Jul07', 'Aug07', 'Sep07', 'Nov08'],
[‘business+web', 'security', 'social+networks']": ['Nov07', 'Jan08', 'Jan09', 'Feb09', 'Aug09']}

hosting and customization (9, 5) - {
[‘scalability', 'social+networks']": ['Nov05'],
[‘business+processes']": ['Apr07'],
[‘dynamic+allocation', 'productivity']": ['Oct06'],
[‘time+to+market', 'business+web']": ['Sep09'],
[‘productivity', 'dynamic+allocation']": ['Dec06'],
[‘open+standards', 'business+processes']": ['Jun08'],
[‘social+networks']": ['Jan09'],
[‘cost']": ['Oct07'],
[‘open+standards', 'privacy', 'time+to+market']": ['Mar09'],
[‘security', 'device+access']": ['Feb07'],
[‘cost', 'privacy']": ['Jul09'],
[‘customer+relationship+management']": ['Aug08', 'Sep08'],
[‘scalability', 'dynamic+allocation', 'managed+services']": ['Nov06'],
[‘productivity', 'managed+services']": ['Oct08'],
[‘social+networks', 'open+standards', 'time+to+market']": ['May08'],
[‘privacy', 'cost']": ['Apr09'],
[‘managed+services', 'productivity']": ['May09'],
[‘movement+of+applications', 'dynamic+allocation', 'outsourcing']": ['Nov06'],
[‘long-term+commitments', 'global+marketplace', 'outsourcing']": ['Feb06']}

Appendix 2 Kenya broadband term clusters

Broadband subsystem

skilled+labor and connectivity+choice (28, 7)

GPD+per+capita and BPO+policy (17, 4) - {
“[‘electricity+distribution+grid', 'BPO']": ['Feb07'],
“[‘BPO', 'aggregate+broadband+demand']": ['Jan09'],
‘[]': ['Feb07', 'Mar08', 'Mar08', 'Apr08', 'Apr08', 'Apr08', 'Jun08', 'Nov08', 'Nov08', 'Jan09', 'Jan09', 'Aug09', 'Aug09', 'Aug09', 'Aug09', 'Oct09', 'Oct09'],
“[‘ICT+equipment+importation+policy']": ['Jun08'],
“[‘ICT+equipment+importation+policy', 'aggregate+broadband+demand']": ['Apr08'],
“[‘BPO', 'ICT+manufacture+maintenance']": ['Oct09'],
“[‘aggregate+broadband+demand', 'ICT+equipment+importation+policy']": ['Apr08'],
“[‘broadband+connected+terminals', 'skilled+labor', 'e-government']": ['Aug09'],
“[‘ICT+sector+investment+competition+policy', 'ICT+equipment+importation+policy', 'macro-economic+factors']": ['Aug09'],
“[‘aggregate+broadband+demand']": ['Mar08'],
“[‘ICT+manufacture+maintenance', 'ICT+equipment+importation+policy', 'electricity+distribution+grid']": ['Oct09']}

e-government+policy and e-government (11, 10) - {
“[‘broadband+infrastructure', 'research+development', 'economic+activity']": ['Sep09'],
“[‘research+development', 'e-security']": ['Mar09'],
“[‘electricity+distribution+grid', 'skilled+labor']": ['Sep09'],
“[‘e-security', 'research+development']": ['Jan09'],
“[‘price+broadband', 'e-health']": ['Nov09'],
“[‘electricity+distribution+grid']": ['Nov09'],
“[‘price+broadband']": ['Dec09'],
“[‘skilled+labor', 'connectivity+choice']": ['Aug09'],
‘[]': ['Apr08', 'Jan09', 'Mar09', 'Mar09', 'Apr09', 'Apr09', 'Jul09', 'Jul09', 'Aug09', 'Sep09', 'Sep09', 'Sep09', 'Nov09', 'Nov09', 'Nov09', 'Dec09', 'Dec09'],
“[‘broadband+infrastructure', 'research+development']": ['Apr09'],
“[‘research+development']": ['Jul09'],
“[‘accessibility', 'research+development']": ['Apr09'],
“[‘research+development', 'accessibility']": ['Jul09'],
“[‘research+development', 'accessibility', 'e-security']": ['Sep09'],
“[‘skilled+labor', 'local+content+services+products', 'electricity+distribution+grid']": ['Mar09'],
“[‘macro-economic+factors']": ['Apr08'],
“[‘e-health', 'price+broadband']": ['Nov09']}

ICT+manufacture+maintenance and GPD+per+capita (20, 17) - {

“[‘ICT+equipment+importation+policy', 'electricity+distribution+grid', 'BPO+policy']":
['Oct09'],
“[‘BPO', 'BPO+policy']": ['Oct09'],
“[‘ICT+equipment+importation+policy', 'electricity+distribution+grid']": ['Feb09'],
“[‘electricity+distribution+grid', 'ICT+equipment+importation+policy']": ['Nov09'],
‘[]': ['Feb07', 'Jun07', 'Sep08', 'Sep08', 'Feb09', 'Feb09', 'Apr09', 'Jul09', 'Oct09', 'Oct09', 'Nov09',
'Nov09', 'Dec09', 'Dec09', 'Dec09'],
“[‘skilled+labor', 'accessibility']": ['Dec09'],
“[‘aggregate+broadband+demand']": ['Jun07', 'Apr09', 'Jul09']}

price+broadband and e-health (25, 12) - {
“[‘e-government', 'e-government+policy']": ['Nov09'],
“[‘e-government+policy', 'e-government']": ['Nov09'],
“[‘accessibility', 'e-security', 'e-government']": ['Jun07'],
‘[]': ['Jun07', 'Sep08', 'Sep08', 'Jan09', 'Jan09', 'Mar09', 'Mar09', 'Apr09', 'Apr09', 'Aug09',
'Aug09', 'Sep09', 'Sep09', 'Nov09', 'Nov09', 'Nov09']}

skilled+labor and connectivity+choice (28, 7) - {
“[‘broadband+connected+terminals', 'local+foreign+investment+policy']": ['Aug07'],
“[‘skilled+labor']": ['Aug09', 'Aug09', 'Nov09', 'Nov09'],
“[‘GPD+per+capita', 'aggregate+broadband+demand', 'e-government+policy']": ['Oct07'],
“[‘e-government', 'e-government+policy']": ['Aug09'],
‘[]': ['Jun07', 'Jun07', 'Aug07', 'Aug07', 'Oct07', 'Oct07', 'Sep08', 'Sep08', 'Nov08', 'Feb09',
'Mar09', 'Aug09', 'Aug09', 'Aug09', 'Nov09', 'Nov09'],
“[‘e-health', 'e-security']": ['Feb09'],
“[‘broadband+infrastructure', 'research+development']": ['Nov08'],
“[‘e-government+policy']": ['Sep08'],
“[‘ICT+sector+investment+competition+policy']": ['Mar09'],
“[‘e-learning']": ['Jun07', 'Jun07']}

Socio-economic subsystem

product+process+innovation and innovation+policy (15, 9) - {
“[‘visibility+communication+opportunities']": ['Feb09'],
“[‘labor+policy']": ['Jul07'],
“[‘income', 'market+transparency', 'industry+efficiency']": ['Feb07'],
“[‘collaboration+knowledge+sharing', 'labor+policy', 'anti-trust+industry+policy']": ['Jul08'],
“[‘employment+opportunities']": ['Jun07'],
“[‘anti-trust+industry+policy', 'transaction+distribution+costs']": ['Jul07'],
“[‘local+foreign+investment+policy', 'employment+opportunities']": ['Dec07'],
“[‘economic+activity', 'market+efficiency', 'local+foreign+investment+policy']": ['Nov09'],
“[‘industry+efficiency', 'labor+policy', 'productivity']": ['May06'],
‘[]': ['May06', 'Feb07', 'Jun07', 'Jul07', 'Jul07', 'Nov07', 'Dec07', 'Mar08', 'Jul08', 'Jul08', 'Jan09',
'Feb09', 'Feb09', 'May09', 'Nov09'],
“[‘employment+opportunities', 'labor+policy']": ['Mar08'],

“[‘collaboration+knowledge+sharing']": ['May09']}

local+foreign+investment+policy and innovation+policy (12, 9) - {
“[‘market+efficiency', 'anti-trust+industry+policy']": ['Aug09'],
“[‘economic+activity', 'market+efficiency', 'product+process+innovation']": ['Nov09'],
“[‘productivity', 'economic+activity', 'income']": ['Aug06'],
“[‘employment+opportunities', 'ICT+manufacture+maintenance', 'market+efficiency']":
['Aug08'],
“[‘social+impact']": ['Oct07'],
‘[]': ['Apr05', 'Apr05', 'Apr06', 'Jun06', 'Aug06', 'Jun07', 'Jun07', 'Oct07', 'Dec07', 'Apr08',
'Aug08', 'Sep08', 'May09', 'Aug09', 'Nov09'],
“[‘transaction+distribution+costs', 'anti-trust+industry+policy', 'employment+opportunities']":
['Sep08'],
“[‘social+service+availability', 'social+impact']": ['Apr06'],
“[‘e-services+products', 'anti-trust+industry+policy']": ['Apr05'],
“[‘product+process+innovation', 'employment+opportunities']": ['Dec07'],
“[‘collaboration+knowledge+sharing', 'visibility+communication+opportunities']": ['Jun07']}

social+service+availability and social+impact (19, 18) - {
“[‘industry+efficiency', 'economic+activity']": ['Jun06'],
“[‘e-services+products', 'jobs', 'employment+opportunities']": ['Feb08'],
“[‘jobs', 'local+foreign+investment+policy']": ['Apr07'],
“[‘jobs']": ['Feb08', 'May08', 'Sep08'],
‘[]': ['Apr06', 'Jun06', 'Jun06', 'Feb07', 'Feb07', 'Apr07', 'Jul07', 'Feb08', 'Feb08', 'Mar08', 'Mar08',
'May08', 'Jun08', 'Aug08', 'Sep08', 'Oct08', 'Feb09', 'Feb09', 'Jun09', 'Oct09', 'Oct09'],
“[‘productivity', 'labor+policy']": ['Oct09'],
“[‘industry+efficiency', 'jobs']": ['Feb09', 'Oct09'],
“[‘innovation+policy', 'local+foreign+investment+policy']": ['Apr06'],
“[‘local+foreign+investment+policy', 'e-services+products']": ['Feb07'],
“[‘economic+activity', 'income']": ['Feb09']}

Education subsystem

research+development and digital+education+policy (14, 4)
research+development and basic+education (14, 1)
certification+policy and certification+examination (3, 2)
innovation+policy and certification+policy (9, 3)
education+investment+policy and basic+education (5, 1)
digital+education+policy and basic+education (4, 1)
education+investment+policy and digital+education+policy (5, 4)
local+content+services+products and formal+training (12, 7)
research+development and e-learning (14, 6)
population and e-learning (13, 6)
e-learning and basic+education (6, 1)
learning+mechanism and certification+policy (11, 3)

research+development and local+content+services+products (14, 12)
informal+training and formal+training (8, 7)
e-learning and education+investment+policy (6, 5)
skills+literacy+level and informal+training (16, 8)
institutional+learning and e-learning (10, 6)
research+development and population (14, 13)
population and basic+education (13, 1)

e-learning and basic+education (6, 1) - {
“[‘research+development', 'education+investment+policy', 'population']": ['May08'],
“[‘local+content+services+products']": ['Sep09'],
“[‘research+development', 'education+investment+policy']": ['Feb09'],
“[‘local+content+services+products', 'learning+mechanism', 'digital+education+policy']":
['Jan08'],
“[‘digital+education+policy']": ['Oct08', 'Aug09'],
“[‘skilled+labor']": ['Apr09'],
“[‘certification+examination', 'certification+policy', 'innovation+policy']": ['Oct09'],
‘[]': ['Feb05', 'Jun05', 'May06', 'Mar07', 'Jun07', 'Jun07', 'Dec07', 'Jan08', 'May08', 'May08',
'Sep08', 'Oct08', 'Feb09', 'Apr09', 'Aug09', 'Sep09', 'Oct09'],
“[‘skills+literacy+level']": ['May06'],
“[‘skills+literacy+level', 'certification+policy']": ['Jun07'],
“[‘education+investment+policy', 'population', 'skilled+labor']": ['Mar07'],
“[‘population', 'research+development', 'local+content+services+products']": ['Jun07'],
“[‘institutional+learning', 'learning+mechanism', 'certification+policy']": ['Dec07'],
“[‘formal+training', 'research+development']": ['Feb05'],
“[‘certification+examination', 'learning+mechanism']": ['Sep08']}

informal+training and formal+training (8, 7) - {
“[‘skilled+labor', 'learning+mechanism', 'innovation+policy']": ['Jan07'],
“[‘education+investment+policy']": ['Sep05'],
“[‘basic+education']": ['Jun07'],
“[‘population']": ['Apr08'],
‘[]': ['Sep05', 'Sep05', 'Sep06', 'Oct06', 'Jan07', 'Jun07', 'Jun07', 'Aug07', 'Feb08', 'Apr08', 'Apr08',
'Aug08', 'Sep08', 'Mar09', 'Jun09', 'Jun09', 'Aug09', 'Oct09', 'Dec09'],
“[‘skills+literacy+level']": ['Sep08'],
“[‘research+development', 'skilled+labor']": ['Feb08'],
“[‘education+investment+policy', 'skilled+labor']": ['Aug07'],
“[‘institutional+learning', 'education+investment+policy']": ['Jun07'],
“[‘certification+policy']": ['Jun09'],
“[‘local+content+services+products', 'skills+literacy+level']": ['Sep06'],
“[‘innovation+policy', 'certification+policy', 'certification+examination']": ['Dec09'],
“[‘institutional+learning']": ['Mar09']}

institutional+learning and e-learning (10, 6) - {
“[‘certification+examination', 'research+development']": ['Dec05'],

“[‘education+investment+policy']": ['Sep06'],
“[‘digital+education+policy']": ['Jun06'],
“[‘skilled+labor', 'certification+examination', 'certification+policy']": ['Apr08'],
‘[]': ['Feb05', 'Dec05', 'Dec05', 'Dec05', 'Feb06', 'Feb06', 'May06', 'Jun06', 'Sep06', 'Sep06',
'Oct06', 'Oct06', 'Feb07', 'Jul07', 'Aug07', 'Oct07', 'Dec07', 'Dec07', 'Mar08', 'Mar08', 'Apr08',
'Apr08', 'Jun08', 'Feb09', 'May09', 'Nov09', 'Dec09'],
“[‘learning+mechanism']": ['Dec07'],
“[‘skills+literacy+level']": ['May06', 'Feb07', 'Jun08'],
“[‘research+development']": ['Dec05'],
“[‘learning+mechanism', 'informal+training', 'skilled+labor']": ['Jul07'],
“[‘population', 'certification+examination']": ['Sep06'],
“[‘certification+examination']": ['Dec05'],
“[‘skilled+labor']": ['Feb09'],
“[‘population', 'innovation+policy']": ['Mar08'],
“[‘basic+education', 'learning+mechanism', 'certification+policy']": ['Dec07']}

research+development and population (14, 13) - {
“[‘basic+education', 'e-learning', 'local+content+services+products']": ['Jun07'],
“[‘local+content+services+products']": ['Mar08'],
“[‘digital+education+policy']": ['Jan07', 'Jun09'],
“[‘basic+education', 'digital+education+policy', 'education+investment+policy']": ['Oct09'],
“[‘basic+education']": ['Apr08', 'Jul08'],
‘[]': ['Aug06', 'Oct06', 'Jan07', 'Jun07', 'Aug07', 'Mar08', 'Apr08', 'May08', 'Jul08', 'Jul08', 'Jul08',
'Aug08', 'Feb09', 'Feb09', 'Jun09', 'Aug09', 'Sep09', 'Oct09', 'Nov09'],
“[‘education+investment+policy', 'basic+education', 'e-learning']": ['May08'],
“[‘digital+education+policy', 'basic+education']": ['Jul08'],
“[‘e-learning']": ['Aug06', 'Feb09'],
“[‘innovation+policy', 'skills+literacy+level']": ['Aug07']}

skilled+labor and informal+training (15, 8) - {
“[‘institutional+learning', 'learning+mechanism', 'e-learning']": ['Jul07'],
‘[]': ['Feb06', 'Feb06', 'Sep06', 'Jan07', 'Feb07', 'Feb07', 'Jul07', 'Jul07', 'Aug07', 'Dec07', 'Dec07',
'Feb08', 'Feb09', 'Feb09', 'Nov09'],
“[‘certification+examination', 'local+content+services+products']": ['Feb09'],
“[‘formal+training', 'learning+mechanism', 'innovation+policy']": ['Jan07'],
“[‘certification+policy']": ['Feb07'],
“[‘certification+examination', 'population', 'education+investment+policy']": ['Feb07'],
“[‘formal+training', 'education+investment+policy']": ['Aug07'],
“[‘formal+training', 'research+development']": ['Feb08'],
“[‘innovation+policy', 'skills+literacy+level']": ['Jul07'],
“[‘innovation+policy', 'certification+policy', 'skills+literacy+level']": ['Nov09'],
“[‘learning+mechanism', 'certification+policy', 'digital+education+policy']": ['Sep06']}

population and basic+education (13, 1) - {
“[‘education+investment+policy']": ['Apr08', 'Sep08'],

“[‘digital+education+policy']": ['Jan07'],
“[‘research+development', 'education+investment+policy', 'e-learning']": ['May08'],
“[‘education+investment+policy', 'local+content+services+products']": ['Aug09'],
‘[]': ['Sep05', 'Sep05', 'Feb06', 'Oct06', 'Jan07', 'Mar07', 'Jun07', 'Apr08', 'Apr08', 'May08', 'Jul08', 'Jul08', 'Sep08', 'Aug09', 'Oct09'],
“[‘research+development']": ['Apr08', 'Jul08'],
“[‘research+development', 'digital+education+policy']": ['Jul08'],
“[‘research+development', 'e-learning', 'local+content+services+products']": ['Jun07'],
“[‘certification+examination']": ['Sep05'],
“[‘learning+mechanism', 'informal+training', 'certification+examination']": ['Sep05'],
“[‘education+investment+policy', 'e-learning', 'skilled+labor']": ['Mar07'],
“[‘research+development', 'digital+education+policy', 'education+investment+policy']": ['Oct09']}

Electricity subsystem

total+electricty+produced and electricity+consumption (16, 6)
population and electricity+consumption (12, 6)
hydroelectricty and geothermal+electricty (10, 9)
electricity+consumption and commercial+industrial+electricity+consumption (6, 4)
population+distribution and population (13, 12)
total+electricty+produced and non-renewable+energy+production (16, 11)
renewable+energy+production and non-renewable+energy+production (14, 11)

total+electricty+produced and electricity+consumption (16, 6) - {
“[‘non-renewable+energy+production', 'energy+policy', 'commercial+industrial+electricity+consumption']": ['Jun06'],
“[‘population', 'population+distribution', 'electricity+distribution+grid']": ['Apr08'],
“[‘commercial+industrial+electricity+consumption']": ['May09'],
“[‘renewable+energy+production', 'population+distribution', 'population']": ['Aug09'],
“[‘commercial+industrial+electricity+consumption', 'renewable+energy+production']": ['Oct06'],
“[‘non-renewable+energy+production', 'population', 'geothermal+electricty']": ['Jun07'],
“[‘population', 'population+distribution']": ['Apr06', 'Nov08'],
‘[]': ['Apr06', 'Jun06', 'Jul06', 'Oct06', 'Jun07', 'Jul07', 'Aug07', 'Apr08', 'Apr08', 'May08', 'Jun08', 'Nov08', 'May09', 'May09', 'Aug09'],
“[‘broadband+infrastructure', 'electricity+distribution+grid', 'residential+electricity+consumption']": ['May09'],
“[‘aggregate+electricity+demand', 'non-renewable+energy+production', 'residential+electricity+consumption']": ['Aug07'],
“[‘economic+activity', 'renewable+energy+production']": ['Jul06'],
“[‘broadband+infrastructure', 'capacity+gap', 'population']": ['Jun08'],
“[‘economic+activity', 'capacity+gap']": ['Jul07']}

population and electricity+consumption (12, 6) - {
“[‘broadband+infrastructure', 'population+distribution']": ['Aug08'],

“[‘population+distribution', 'economic+activity',
'commercial+industrial+electricity+consumption']": ['Nov05'],
“[‘non-renewable+energy+production', 'geothermal+electricty', 'total+electricty+produced']":
['Jun07'],
“[‘energy+policy', 'population+distribution']": ['Aug06'],
“[‘energy+policy']": ['Aug09'],
“[‘population+distribution']": ['Mar08'],
‘[]': ['Oct05', 'Nov05', 'Apr06', 'Aug06', 'Apr07', 'Jun07', 'Mar08', 'Apr08', 'Jun08', 'Jun08',
'Aug08', 'Oct08', 'Nov08', 'Nov08', 'Aug09', 'Aug09', 'Aug09'],
“[‘renewable+energy+production', 'total+electricty+produced', 'population+distribution']":
['Aug09'],
“[‘energy+policy', 'renewable+energy+production']": ['Aug09'],
“[‘residential+electricity+consumption', 'broadband+infrastructure']": ['Apr07'],
“[‘broadband+infrastructure', 'energy+policy', 'economic+activity']": ['Nov08'],
“[‘total+electricty+produced', 'population+distribution', 'electricity+distribution+grid']":
['Apr08'],
“[‘broadband+infrastructure', 'total+electricty+produced', 'capacity+gap']": ['Jun08'],
“[‘economic+activity', 'population+distribution', 'renewable+energy+production']": ['Oct08'],
“[‘residential+electricity+consumption', 'population+distribution']": ['Jun08'],
“[‘total+electricty+produced', 'population+distribution']": ['Apr06', 'Nov08']}

hydroelectricty and geothermal+electricty (10, 9) - {
“[‘electricity+distribution+grid', 'transmission+mechanism']": ['Aug08'],
“[‘transmission+mechanism', 'electricity+distribution+grid', 'renewable+energy+production']":
['Sep09'],
“[‘aggregate+electricity+demand', 'electricity+distribution+grid']": ['Nov07'],
“[‘economic+activity', 'electricity+distribution+grid', 'transmission+mechanism']": ['Feb06'],
“[‘broadband+infrastructure', 'commercial+industrial+electricity+consumption',
'total+electricty+produced']": ['Jan09'],
“[‘residential+electricity+consumption', 'transmission+mechanism', 'non-
renewable+energy+production']": ['Feb07'],
“[‘non-renewable+energy+production', 'aggregate+electricity+demand',
'electricity+distribution+grid']": ['Apr06'],
“[‘commercial+industrial+electricity+consumption', 'electricity+consumption',
'electricity+distribution+grid']": ['Feb07'],
“[‘electricity+distribution+grid']": ['Jul07', 'Jun08'],
“[‘aggregate+electricity+demand']": ['Jun07', 'Mar08', 'Nov09'],
‘[]': ['Oct05', 'Oct05', 'Oct05', 'Feb06', 'Feb06', 'Apr06', 'May06', 'May06', 'Jul06', 'Jul06', 'Dec06',
'Feb07', 'Feb07', 'Jun07', 'Jul07', 'Jul07', 'Sep07', 'Oct07', 'Oct07', 'Nov07', 'Nov07', 'Jan08',
'Feb08', 'Mar08', 'Mar08', 'Jun08', 'Jun08', 'Aug08', 'Aug08', 'Aug08', 'Jan09', 'Jan09', 'May09',
'Jul09', 'Sep09', 'Sep09', 'Oct09', 'Oct09', 'Nov09'],
“[‘residential+electricity+consumption']": ['Oct07'],
“[‘residential+electricity+consumption', 'renewable+energy+production']": ['Oct05'],
“[‘residential+electricity+consumption', 'aggregate+electricity+demand']": ['Jan09'],
“[‘transmission+mechanism']": ['Jul06', 'Dec06', 'Oct07', 'Jul09', 'Oct09'],

“[‘commercial+industrial+electricity+consumption', 'residential+electricity+consumption', 'renewable+energy+production']": ['Jan08'],
“[‘non-renewable+energy+production', 'capacity+gap']": ['Jul06'],
“[‘commercial+industrial+electricity+consumption', 'electricity+distribution+grid', 'non-renewable+energy+production']": ['Aug08'],
“[‘transmission+mechanism', 'aggregate+electricity+demand']": ['Oct05'],
“[‘energy+policy']": ['Oct05'],
“[‘aggregate+electricity+demand', 'transmission+mechanism', 'electricity+distribution+grid']": ['May09']}

geothermal+electricty and electricity+distribution+grid (9, 7) - {
“[‘aggregate+electricity+demand', 'hydroelectricty', 'transmission+mechanism']": ['May09'],
“[‘total+electricty+produced', 'capacity+gap', 'non-renewable+energy+production']": ['Jan08'],
“[‘commercial+industrial+electricity+consumption', 'electricity+consumption', 'hydroelectricty']": ['Feb07'],
“[‘hydroelectricty', 'transmission+mechanism']": ['Aug08'],
“[‘economic+activity', 'transmission+mechanism', 'hydroelectricty']": ['Feb06'],
“[‘transmission+mechanism', 'residential+electricity+consumption']": ['Oct09'],
“[‘aggregate+electricity+demand']": ['Oct06', 'Feb08'],
‘[]': ['Feb06', 'Apr06', 'Jun06', 'Oct06', 'Feb07', 'Feb07', 'Jul07', 'Sep07', 'Nov07', 'Jan08', 'Feb08', 'Jun08', 'Aug08', 'Aug08', 'Feb09', 'May09', 'May09', 'Jul09', 'Sep09', 'Oct09'],
“[‘hydroelectricty']": ['Jul07', 'Jun08'],
“[‘residential+electricity+consumption']": ['May09'],
“[‘non-renewable+energy+production', 'total+electricty+produced']": ['Jul09'],
“[‘aggregate+electricity+demand', 'transmission+mechanism']": ['Jun06'],
“[‘transmission+mechanism']": ['Feb09'],
“[‘transmission+mechanism', 'hydroelectricty', 'renewable+energy+production']": ['Sep09'],
“[‘commercial+industrial+electricity+consumption', 'non-renewable+energy+production', 'hydroelectricty']": ['Aug08'],
“[‘non-renewable+energy+production', 'hydroelectricty', 'aggregate+electricity+demand']": ['Apr06'],
“[‘aggregate+electricity+demand', 'hydroelectricty']": ['Nov07']}

transmission+mechanism and aggregate+electricity+demand (17, 1) - {
“[‘geothermal+electricty', 'hydroelectricty']": ['Oct05'],
“[‘hydroelectricty', 'energy+policy']": ['Jul09'],
‘[]': ['Oct05', 'Oct05', 'Feb06', 'Jun06', 'Jul06', 'Aug07', 'May08', 'Nov08', 'Feb09', 'Apr09', 'May09', 'Jul09', 'Sep09', 'Sep09', 'Dec09'],
“[‘hydroelectricty', 'electricity+distribution+grid', 'geothermal+electricty']": ['May09'],
“[‘residential+electricity+consumption', 'renewable+energy+production', 'electricity+consumption']": ['Sep09'],
“[‘geothermal+electricty', 'capacity+gap']": ['Feb06'],
“[‘electricity+distribution+grid', 'geothermal+electricty']": ['Jun06'],
“[‘non-renewable+energy+production']": ['Feb09']}

population+distribution and population (13, 12) - {
“[‘electricity+consumption', 'broadband+infrastructure']": ['Aug08'],
“[‘renewable+energy+production', 'non-renewable+energy+production', 'energy+policy']":
['Dec09'],
“[‘total+electricty+produced', 'electricity+consumption']": ['Apr06', 'Nov08'],
“[‘total+electricty+produced', 'electricity+consumption', 'electricity+distribution+grid']":
['Apr08'],
“[‘residential+electricity+consumption', 'commercial+industrial+electricity+consumption']":
['Oct06', 'Oct06'],
“[‘broadband+infrastructure']": ['Nov09'],
“[‘electricity+distribution+grid']": ['Sep08'],
“[‘economic+activity']": ['Jul09'],
‘[]': ['Nov05', 'Feb06', 'Feb06', 'Apr06', 'May06', 'Jun06', 'Jun06', 'Aug06', 'Aug06', 'Oct06',
'Oct06', 'Apr07', 'Aug07', 'Aug07', 'Nov07', 'Nov07', 'Mar08', 'Apr08', 'May08', 'Jun08', 'Jul08',
'Aug08', 'Aug08', 'Sep08', 'Sep08', 'Oct08', 'Nov08', 'Nov08', 'Jan09', 'May09', 'Jul09', 'Jul09',
'Aug09', 'Oct09', 'Nov09', 'Dec09', 'Dec09'],
“[‘hydroelectricty']": ['Jun06'],
“[‘economic+activity', 'renewable+energy+production', 'electricity+consumption']": ['Oct08'],
“[‘total+electricty+produced']": ['Sep08'],
“[‘energy+policy', 'electricity+consumption']": ['Aug06'],
“[‘residential+electricity+consumption', 'electricity+consumption']": ['Jun08'],
“[‘economic+activity', 'commercial+industrial+electricity+consumption',
'electricity+consumption']": ['Nov05'],
“[‘renewable+energy+production', 'electricity+consumption', 'total+electricty+produced']":
['Aug09'],
“[‘commercial+industrial+electricity+consumption', 'economic+activity', 'energy+policy']":
['Feb06'],
“[‘electricity+consumption']": ['Mar08']}

total+electricty+produced and non-renewable+energy+production (16, 11) - {
“[‘electricity+consumption', 'energy+policy', 'commercial+industrial+electricity+consumption']":
['Jun06'],
“[‘transmission+mechanism', 'commercial+industrial+electricity+consumption']": ['Jan09'],
“[‘capacity+gap', 'geothermal+electricty', 'electricity+distribution+grid']": ['Jan08'],
‘[]': ['Oct05', 'Oct05', 'May06', 'Jun06', 'Jun06', 'Jun06', 'Aug06', 'Dec06', 'Jun07', 'Aug07',
'Dec07', 'Jan08', 'Nov08', 'Dec08', 'Jan09', 'Jul09'],
“[‘electricity+consumption', 'population', 'geothermal+electricty']": ['Jun07'],
“[‘hydroelectricty']": ['Dec08'],
“[‘capacity+gap', 'commercial+industrial+electricity+consumption', 'hydroelectricty']": ['Jun06'],
“[‘capacity+gap', 'energy+policy', 'renewable+energy+production']": ['Jun06'],
“[‘commercial+industrial+electricity+consumption', 'electricity+distribution+grid']": ['Dec07'],
“[‘electricity+distribution+grid', 'population+distribution', 'economic+activity']": ['Oct05'],
“[‘electricity+consumption', 'aggregate+electricity+demand',
'residential+electricity+consumption']": ['Aug07'],
“[‘population+distribution', 'electricity+distribution+grid']": ['Oct05'],

“[‘broadband+infrastructure', 'energy+policy']": ['Nov08'],
“[‘electricity+distribution+grid', 'geothermal+electricty']": ['Jul09'],
“[‘renewable+energy+production']": ['Aug06']}

geothermal+electricty and aggregate+electricity+demand (9, 1) - {
“[‘transmission+mechanism', 'electricity+distribution+grid']": ['Jun06'],
“[‘residential+electricity+consumption', 'hydroelectricty']": ['Jan09'],
“[‘transmission+mechanism', 'capacity+gap']": ['Feb06'],
“[‘electricity+distribution+grid']": ['Oct06', 'Feb08'],
‘[]': ['Oct05', 'Feb06', 'Apr06', 'Jun06', 'Oct06', 'Jun07', 'Jun07', 'Sep07', 'Nov07', 'Nov07', 'Feb08', 'Mar08', 'Jan09', 'Feb09', 'Mar09', 'Mar09', 'Apr09', 'May09', 'May09', 'Aug09', 'Aug09', 'Nov09'],
“[‘hydroelectricty']": ['Jun07', 'Mar08', 'Nov09'],
“[‘hydroelectricty', 'electricity+distribution+grid']": ['Nov07'],
“[‘hydroelectricty', 'transmission+mechanism', 'electricity+distribution+grid']": ['May09'],
“[‘residential+electricity+consumption', 'electricity+consumption']": ['Jun07'],
“[‘non-renewable+energy+production', 'hydroelectricty', 'electricity+distribution+grid']": ['Apr06'],
“[‘transmission+mechanism', 'hydroelectricty']": ['Oct05']}

hydroelectricty and aggregate+electricity+demand (10, 1) - {
“[‘economic+activity', 'capacity+gap', 'residential+electricity+consumption']": ['Mar09'],
“[‘commercial+industrial+electricity+consumption']": ['Aug07'],
“[‘electricity+distribution+grid']": ['Apr08'],
“[‘transmission+mechanism', 'electricity+distribution+grid', 'geothermal+electricty']": ['May09'],
“[‘geothermal+electricty', 'electricity+distribution+grid']": ['Nov07'],
“[‘geothermal+electricty', 'residential+electricity+consumption']": ['Jan09'],
“[‘transmission+mechanism', 'geothermal+electricty']": ['Oct05'],
‘[]': ['Oct05', 'Apr06', 'Jun07', 'Jun07', 'Aug07', 'Nov07', 'Mar08', 'Apr08', 'Apr08', 'May08', 'Jul08', 'Jan09', 'Mar09', 'May09', 'Jul09', 'Jul09', 'Nov09'],
“[‘geothermal+electricty']": ['Jun07', 'Mar08', 'Nov09'],
“[‘transmission+mechanism', 'energy+policy']": ['Jul09'],
“[‘non-renewable+energy+production', 'electricity+distribution+grid', 'geothermal+electricty']": ['Apr06']}

renewable+energy+production and non-renewable+energy+production (14, 11) - {
“[‘commercial+industrial+electricity+consumption']": ['Oct09'],
“[‘population+distribution', 'population', 'energy+policy']": ['Dec09'],
“[‘electricity+distribution+grid']": ['Jul09'],
‘[]': ['Dec05', 'Feb06', 'Jun06', 'Aug06', 'Aug06', 'Dec06', 'Aug07', 'Oct07', 'Feb08', 'Apr08', 'Jun08', 'Aug08', 'Dec08', 'May09', 'Jul09', 'Oct09', 'Dec09'],
“[‘total+electricty+produced']": ['Aug06'],
“[‘aggregate+electricity+demand', 'residential+electricity+consumption']": ['Aug06'],
“[‘energy+policy', 'residential+electricity+consumption']": ['May09'],
“[‘capacity+gap', 'total+electricty+produced', 'energy+policy']": ['Jun06'],
“[‘electricity+consumption']": ['Dec05']}

transmission+mechanism and geothermal+electricty (17, 9) - {
“[‘residential+electricity+consumption', 'hydroelectricty', 'non-renewable+energy+production']": ['Feb07'],
“[‘electricity+distribution+grid', 'hydroelectricty', 'renewable+energy+production']": ['Sep09'],
“[‘electricity+distribution+grid']": ['Feb09'],
‘[]': ['Oct05', 'Feb06', 'Feb06', 'Jun06', 'Jul06', 'Dec06', 'Dec06', 'Feb07', 'Oct07', 'Dec07', 'Apr08', 'Jul08', 'Aug08', 'Nov08', 'Feb09', 'May09', 'Jul09', 'Sep09', 'Oct09', 'Oct09'],
“[‘hydroelectricty']": ['Jul06', 'Dec06', 'Oct07', 'Jul09', 'Oct09'],
“[‘electricity+distribution+grid', 'residential+electricity+consumption']": ['Oct09'],
“[‘aggregate+electricity+demand', 'hydroelectricty', 'electricity+distribution+grid']": ['May09'],
“[‘economic+activity', 'electricity+distribution+grid', 'hydroelectricty']": ['Feb06'],
“[‘broadband+infrastructure']": ['Jul08'],
“[‘aggregate+electricity+demand', 'electricity+distribution+grid']": ['Jun06'],
“[‘aggregate+electricity+demand', 'hydroelectricty']": ['Oct05'],
“[‘hydroelectricty', 'electricity+distribution+grid']": ['Aug08'],
“[‘capacity+gap', 'aggregate+electricity+demand']": ['Feb06']}

National Security and Administration Subsystem
problem+resolution and appropriate+investment+climate
problem+communication and appropriate+investment+climate (9, 1)
broadband+infrastructure and appropriate+investment+climate (2, 1)
national+security+policy and broadband+infrastructure (6, 2)
problem+communication and e-services+products (9, 4)
freedom+information+policy and appropriate+investment+climate (5, 1)
population and broadband+infrastructure (7, 2)
privacy+policy and appropriate+investment+climate (8, 1)
population and national+security+policy (7, 6)
problem+resolution and national+security+policy (12, 6)
problem+resolution and problem+communication (12, 9)
e-services+products and appropriate+investment+climate (4, 1)
problem+communication and population (9, 7)
resolution+mechanism and problem+identification+mechanism (13, 10)
national+security+policy and e-services+products (6, 4)
problem+resolution and population (12, 7)
e-services+products and broadband+infrastructure (4, 2)
problem+communication and national+security+policy (9, 6)
national+security+policy and freedom+information+policy (6, 5)
problem+resolution and broadband+infrastructure (12, 2)
problem+communication and broadband+infrastructure (9, 2)
national+security+policy and appropriate+investment+climate (6, 1)
population and e-services+products (7, 4)
problem+resolution and e-services+products (12, 4)
problem+communication and freedom+information+policy (9, 5)

problem+communication and e-services+products (9, 4) - {
“[‘problem+occurence', 'civic+stability']": ['Feb08'],
“[‘problem+resolution', 'population']": ['Dec07', 'Dec07'],
“[‘broadband+infrastructure', 'national+security+policy', 'problem+resolution']": ['Mar09'],
“[‘broadband+infrastructure', 'national+security+policy', 'population']": ['Feb06'],
“[‘privacy+policy', 'problem+resolution', 'broadband+infrastructure']": ['Oct08'],
“[‘resolution+mechanism']": ['Mar09'],
“[‘problem+resolution', 'national+security+policy', 'broadband+infrastructure']": ['Aug07'],
“[‘population']": ['Dec09'],
“[‘national+security+policy']": ['Apr06', 'Oct06', 'Dec09'],
“[‘population', 'broadband+infrastructure', 'national+security+policy']": ['Apr06'],
“[‘broadband+infrastructure', 'problem+resolution']": ['May06'],
“[‘problem+occurence']": ['Jan08'],
“[‘population', 'national+security+policy', 'appropriate+investment+climate']": ['Jun08'],
‘[]': ['Feb06', 'Apr06', 'Apr06', 'May06', 'May06', 'Oct06', 'Nov06', 'Dec06', 'Dec06', 'Aug07',
'Dec07', 'Dec07', 'Jan08', 'Feb08', 'Apr08', 'Jun08', 'Jun08', 'Sep08', 'Oct08', 'Dec08', 'Mar09',
'Mar09', 'May09', 'Nov09', 'Nov09', 'Dec09', 'Dec09'],
“[‘freedom+information+policy', 'population']": ['Apr08'],
“[‘broadband+infrastructure', 'population', 'national+security+policy']": ['Dec06'],
“[‘freedom+information+policy']": ['Nov09'],
“[‘population', 'broadband+infrastructure']": ['May06']}

freedom+information+policy and appropriate+investment+climate (5, 1) - {
“[‘national+security+policy', 'e-services+products']": ['May06'],
“[‘e-services+products']": ['Oct07'],
‘[]': ['May06', 'May06', 'Jun06', 'Jul07', 'Oct07', 'Dec07', 'May08', 'Oct08', 'Nov08', 'Nov08',
'Mar09', 'Apr09', 'Apr09', 'Sep09', 'Dec09'],
“[‘problem+identification+mechanism', 'broadband+infrastructure']": ['Dec07'],
“[‘problem+occurence']": ['Apr09'],
“[‘problem+identification+mechanism', 'civic+stability']": ['May08'],
“[‘population', 'broadband+infrastructure', 'problem+resolution']": ['Jun06'],
“[‘privacy+policy']": ['Oct08'],
“[‘resolution+mechanism', 'problem+identification+mechanism', 'e-services+products']":
['Apr09'],
“[‘national+security+policy', 'e-services+products', 'broadband+infrastructure']": ['Nov08']}

population and broadband+infrastructure (7, 2) - {
“[‘problem+resolution', 'appropriate+investment+climate', 'privacy+policy']": ['Jun06'],
“[‘e-services+products', 'problem+communication']": ['May06'],
“[‘problem+resolution', 'problem+communication']": ['Mar07'],
“[‘resolution+mechanism']": ['Dec06'],
“[‘problem+resolution', 'appropriate+investment+climate', 'freedom+information+policy']":
['Jun06'],
“[‘e-services+products', 'national+security+policy', 'problem+communication']": ['Feb06'],

‘[]': ['Feb06', 'Apr06', 'Apr06', 'May06', 'Jun06', 'Jun06', 'Oct06', 'Nov06', 'Dec06', 'Dec06', 'Mar07', 'Nov07', 'Nov07', 'Feb08', 'Aug08', 'Aug08', 'May09', 'Jun09', 'Sep09', 'Nov09'],
“[‘problem+communication', 'e-services+products', 'national+security+policy']": ['Apr06', 'Dec06'],
“[‘national+security+policy']": ['Aug08', 'Sep09'],
“[‘national+security+policy', 'problem+resolution']": ['Feb08'],
“[‘privacy+policy']": ['Oct06'],
“[‘problem+resolution', 'national+security+policy']": ['Nov06'],
“[‘problem+resolution', 'appropriate+investment+climate', 'e-services+products']": ['Nov09']}

problem+occurence and problem+identification+mechanism (11, 10) - {
“[‘resolution+mechanism', 'civic+stability', 'appropriate+investment+climate']": ['Apr07'],
“[‘privacy+policy', 'resolution+mechanism']": ['Jan09'],
“[‘e-services+products']": ['Sep09'],
“[‘resolution+mechanism']": ['Apr06', 'Jun08'],
‘[]': ['Aug05', 'Apr06', 'Jun06', 'Nov06', 'Apr07', 'Apr07', 'Jul07', 'Jan08', 'Feb08', 'Jun08', 'Sep08', 'Nov08', 'Jan09', 'Jun09', 'Jun09', 'Jul09', 'Sep09', 'Sep09', 'Sep09'],
“[‘resolution+mechanism', 'civic+stability']": ['Jul07', 'Jun09', 'Jul09'],
“[‘civic+stability']": ['Sep09'],
“[‘privacy+policy', 'resolution+mechanism', 'e-services+products']": ['Sep09'],
“[‘resolution+mechanism', 'civic+stability', 'freedom+information+policy']": ['Jun06'],
“[‘appropriate+investment+climate', 'e-services+products']": ['Apr07'],
“[‘privacy+policy', 'resolution+mechanism', 'civic+stability']": ['Feb08']}

problem+identification+mechanism and civic+stability (10, 3) - {
“[‘resolution+mechanism', 'appropriate+investment+climate', 'problem+occurence']": ['Apr07'],
“[‘problem+occurence', 'resolution+mechanism', 'freedom+information+policy']": ['Jun06'],
‘[]': ['Apr06', 'Jun06', 'Jan07', 'Mar07', 'Mar07', 'Apr07', 'Apr07', 'Jul07', 'Feb08', 'May08', 'Oct08', 'Mar09', 'Jun09', 'Jul09', 'Sep09'],
“[‘freedom+information+policy', 'appropriate+investment+climate']": ['May08'],
“[‘problem+occurence']": ['Sep09'],
“[‘problem+occurence', 'resolution+mechanism']": ['Jul07', 'Jun09', 'Jul09'],
“[‘resolution+mechanism']": ['Jan07', 'Mar07'],
“[‘resolution+mechanism', 'appropriate+investment+climate']": ['Apr07'],
“[‘freedom+information+policy']": ['Oct08'],
“[‘privacy+policy', 'resolution+mechanism', 'problem+occurence']": ['Feb08'],
“[‘appropriate+investment+climate', 'resolution+mechanism']": ['Apr06']}

population and national+security+policy (7, 6) - {
“[‘e-services+products', 'civic+stability', 'privacy+policy']": ['Aug09'],
“[‘problem+communication', 'problem+resolution']": ['Jul09'],
“[‘broadband+infrastructure', 'problem+communication', 'e-services+products']": ['Dec06'],
“[‘problem+occurence', 'problem+resolution']": ['Jan07'],
“[‘problem+resolution', 'e-services+products']": ['Jun08'],
“[‘e-services+products']": ['Apr07'],

“[‘e-services+products', 'problem+resolution']": ['Jun06'],
‘[]': ['Aug05', 'Feb06', 'Feb06', 'Apr06', 'Jun06', 'Nov06', 'Nov06', 'Dec06', 'Jan07', 'Apr07', 'Oct07', 'Feb08', 'Apr08', 'Jun08', 'Jun08', 'Aug08', 'Jan09', 'Mar09', 'Jun09', 'Jul09', 'Jul09', 'Aug09', 'Sep09', 'Sep09', 'Dec09', 'Dec09'],
“[‘problem+resolution', 'problem+communication', 'freedom+information+policy']": ['Apr08'],
“[‘broadband+infrastructure', 'e-services+products', 'problem+communication']": ['Feb06'],
“[‘problem+communication', 'appropriate+investment+climate', 'problem+resolution']": ['Dec09'],
“[‘broadband+infrastructure']": ['Aug08', 'Sep09'],
“[‘problem+resolution']": ['Mar09'],
“[‘e-services+products', 'problem+communication', 'appropriate+investment+climate']": ['Jun08'],
“[‘problem+communication', 'e-services+products', 'broadband+infrastructure']": ['Apr06'],
“[‘problem+resolution', 'broadband+infrastructure']": ['Nov06', 'Feb08'],
“[‘privacy+policy', 'freedom+information+policy']": ['Feb06']}

problem+resolution and national+security+policy (12, 6) - {
“[‘problem+communication', 'freedom+information+policy', 'population']": ['Apr08'],
“[‘problem+communication', 'broadband+infrastructure', 'e-services+products']": ['Aug07'],
“[‘population', 'problem+occurence']": ['Jan07'],
“[‘e-services+products', 'population']": ['Jun08'],
“[‘broadband+infrastructure', 'population']": ['Nov06'],
“[‘population']": ['Mar09'],
‘[]': ['Jun06', 'Aug06', 'Nov06', 'Jan07', 'Aug07', 'Feb08', 'Apr08', 'Jun08', 'Feb09', 'Mar09', 'Mar09', 'Jun09', 'Jul09', 'Oct09', 'Dec09'],
“[‘problem+communication', 'population', 'appropriate+investment+climate']": ['Dec09'],
“[‘appropriate+investment+climate', 'privacy+policy']": ['Jun09'],
“[‘population', 'e-services+products']": ['Jun06'],
“[‘problem+communication', 'population']": ['Jul09'],
“[‘problem+communication']": ['Oct09'],
“[‘broadband+infrastructure', 'problem+communication', 'e-services+products']": ['Mar09'],
“[‘population', 'broadband+infrastructure']": ['Feb08'],
“[‘problem+communication', 'broadband+infrastructure', 'privacy+policy']": ['Aug06']}

resolution+mechanism and problem+occurence (13, 11) - {
“[‘freedom+information+policy', 'civic+stability']": ['May08'],
“[‘appropriate+investment+climate', 'civic+stability']": ['Oct06'],
‘[]': ['Apr06', 'May06', 'Jun06', 'Oct06', 'Apr07', 'Jul07', 'Feb08', 'May08', 'Jun08', 'Oct08', 'Dec08', 'Dec08', 'Jan09', 'May09', 'Jun09', 'Jul09', 'Sep09', 'Sep09'],
“[‘problem+identification+mechanism', 'civic+stability', 'appropriate+investment+climate']": ['Apr07'],
“[‘privacy+policy', 'problem+identification+mechanism']": ['Jan09'],
“[‘problem+identification+mechanism']": ['Apr06', 'Jun08'],
“[‘problem+identification+mechanism', 'civic+stability', 'freedom+information+policy']": ['Jun06'],

“[‘civic+stability', 'privacy+policy']": ['Sep09'],
“[‘problem+identification+mechanism', 'civic+stability']": ['Jul07', 'Jun09', 'Jul09'],
“[‘civic+stability']": ['Dec08'],
“[‘freedom+information+policy']": ['Dec08'],
“[‘problem+identification+mechanism', 'privacy+policy', 'e-services+products']": ['Sep09'],
“[‘privacy+policy', 'problem+identification+mechanism', 'civic+stability']": ['Feb08']}

problem+resolution and problem+communication (12, 9) - {
“[‘broadband+infrastructure', 'national+security+policy', 'privacy+policy']": ['Aug06'],
“[‘national+security+policy', 'population', 'appropriate+investment+climate']": ['Dec09'],
“[‘appropriate+investment+climate']": ['Jul07'],
“[‘privacy+policy', 'e-services+products', 'broadband+infrastructure']": ['Oct08'],
“[‘privacy+policy', 'problem+identification+mechanism', 'population']": ['Feb07'],
“[‘broadband+infrastructure', 'e-services+products']": ['May06'],
“[‘civic+stability', 'privacy+policy', 'freedom+information+policy']": ['Nov09'],
“[‘broadband+infrastructure']": ['Apr08', 'Jun09'],
“[‘population']": ['Nov05', 'Nov05', 'Jul07'],
‘[]': ['Nov05', 'Nov05', 'May06', 'Aug06', 'Aug06', 'Feb07', 'Mar07', 'Jul07', 'Jul07', 'Aug07',
'Dec07', 'Dec07', 'Feb08', 'Apr08', 'Apr08', 'Oct08', 'Jan09', 'Mar09', 'Jun09', 'Jul09', 'Oct09',
'Nov09', 'Dec09'],
“[‘national+security+policy', 'freedom+information+policy', 'population']": ['Apr08'],
“[‘national+security+policy', 'broadband+infrastructure', 'e-services+products']": ['Aug07'],
“[‘national+security+policy']": ['Oct09'],
“[‘population', 'appropriate+investment+climate']": ['Aug06'],
“[‘population', 'e-services+products']": ['Dec07', 'Dec07'],
“[‘national+security+policy', 'population']": ['Jul09'],
“[‘freedom+information+policy', 'broadband+infrastructure']": ['Feb08'],
“[‘population', 'broadband+infrastructure']": ['Mar07'],
“[‘broadband+infrastructure', 'national+security+policy', 'e-services+products']": ['Mar09']}

problem+communication and population (9, 7) - {
“[‘e-services+products', 'broadband+infrastructure']": ['May06'],
“[‘e-services+products', 'broadband+infrastructure', 'national+security+policy']": ['Apr06'],
“[‘problem+resolution', 'e-services+products']": ['Dec07', 'Dec07'],
“[‘e-services+products']": ['Dec09'],
“[‘e-services+products', 'national+security+policy', 'appropriate+investment+climate']": ['Jun08'],
“[‘appropriate+investment+climate', 'problem+resolution']": ['Aug06'],
‘[]': ['Nov05', 'Nov05', 'Nov05', 'Feb06', 'Apr06', 'May06', 'Aug06', 'Dec06', 'Feb07', 'Mar07',
'Jul07', 'Dec07', 'Dec07', 'Apr08', 'Apr08', 'Jun08', 'Jul09', 'Dec09', 'Dec09'],
“[‘problem+resolution', 'national+security+policy', 'freedom+information+policy']": ['Apr08'],
“[‘broadband+infrastructure', 'e-services+products', 'national+security+policy']": ['Feb06',
'Dec06'],
“[‘e-services+products', 'freedom+information+policy']": ['Apr08'],
“[‘national+security+policy', 'problem+resolution']": ['Jul09'],
“[‘privacy+policy', 'problem+identification+mechanism', 'problem+resolution']": ['Feb07'],

“[‘problem+resolution']": ['Nov05', 'Nov05', 'Jul07'],
“[‘national+security+policy', 'appropriate+investment+climate', 'problem+resolution']":
['Dec09'],
“[‘problem+resolution', 'broadband+infrastructure']": ['Mar07']}

resolution+mechanism and problem+identification+mechanism (13, 10) - {
“[‘problem+occurence', 'civic+stability']": ['Jul07', 'Jun09', 'Jul09'],
“[‘problem+occurence', 'privacy+policy']": ['Jan09'],
“[‘civic+stability', 'appropriate+investment+climate']": ['Apr06', 'Apr07'],
“[‘civic+stability']": ['Jan07', 'Mar07'],
“[‘privacy+policy', 'problem+occurence', 'civic+stability']": ['Feb08'],
“[‘population']": ['May08'],
‘[]': ['Feb06', 'Feb06', 'Apr06', 'Apr06', 'Apr06', 'May06', 'Jun06', 'Aug06', 'Aug06', 'Jan07',
'Jan07', 'Mar07', 'Apr07', 'Apr07', 'Jul07', 'Feb08', 'Mar08', 'May08', 'Jun08', 'Jul08', 'Jan09',
'Jan09', 'Feb09', 'Apr09', 'Apr09', 'May09', 'Jun09', 'Jul09', 'Sep09', 'Oct09', 'Dec09'],
“[‘appropriate+investment+climate', 'national+security+policy', 'privacy+policy']": ['Feb06'],
“[‘problem+occurence']": ['Apr06', 'Jun08'],
“[‘privacy+policy', 'e-services+products', 'problem+occurence']": ['Sep09'],
“[‘problem+occurence', 'civic+stability', 'freedom+information+policy']": ['Jun06'],
“[‘freedom+information+policy']": ['Aug06'],
“[‘civic+stability', 'appropriate+investment+climate', 'problem+occurence']": ['Apr07'],
“[‘freedom+information+policy', 'appropriate+investment+climate', 'e-services+products']":
['Apr09']}

national+security+policy and e-services+products (6, 4) - {
“[‘problem+communication', 'population', 'broadband+infrastructure']": ['Apr06'],
“[‘population', 'problem+communication', 'appropriate+investment+climate']": ['Jun08'],
“[‘civic+stability', 'appropriate+investment+climate']": ['Oct06'],
“[‘appropriate+investment+climate', 'freedom+information+policy']": ['May06'],
“[‘problem+resolution', 'population']": ['Jun08'],
“[‘broadband+infrastructure', 'problem+resolution', 'problem+communication']": ['Mar09'],
“[‘broadband+infrastructure', 'population', 'problem+communication']": ['Feb06'],
‘[]': ['Feb06', 'Apr06', 'Apr06', 'May06', 'Jun06', 'Jun06', 'Aug06', 'Oct06', 'Oct06', 'Dec06',
'Apr07', 'Jul07', 'Aug07', 'Nov07', 'Jun08', 'Jun08', 'Nov08', 'Mar09', 'Aug09', 'Dec09'],
“[‘population']": ['Apr07'],
“[‘civic+stability', 'resolution+mechanism']": ['Jun06'],
“[‘freedom+information+policy', 'broadband+infrastructure', 'appropriate+investment+climate']":
['Nov08'],
“[‘population', 'civic+stability', 'privacy+policy']": ['Aug09'],
“[‘broadband+infrastructure', 'problem+communication', 'population']": ['Dec06'],
“[‘freedom+information+policy']": ['Jul07'],
“[‘problem+communication']": ['Apr06', 'Oct06', 'Dec09'],
“[‘problem+resolution', 'problem+communication', 'broadband+infrastructure']": ['Aug07'],
“[‘population', 'problem+resolution']": ['Jun06']}

problem+resolution and population (12, 7) - {
“[‘problem+communication', 'e-services+products']": ['Dec07', 'Dec07'],
“[‘e-services+products', 'national+security+policy']": ['Jun06'],
“[‘national+security+policy', 'problem+communication', 'freedom+information+policy']":
['Apr08'],
“[‘national+security+policy', 'problem+occurence']": ['Jan07'],
“[‘problem+communication', 'national+security+policy']": ['Jul09'],
“[‘national+security+policy', 'e-services+products']": ['Jun08'],
“[‘national+security+policy', 'broadband+infrastructure']": ['Feb08'],
“[‘problem+communication', 'broadband+infrastructure']": ['Mar07'],
“[‘appropriate+investment+climate', 'problem+communication']": ['Aug06'],
“[‘appropriate+investment+climate', 'privacy+policy', 'broadband+infrastructure']": ['Jun06'],
‘[]': ['Nov05', 'Nov05', 'May06', 'Jun06', 'Jun06', 'Jun06', 'Aug06', 'Nov06', 'Jan07', 'Feb07',
'Mar07', 'Apr07', 'Jul07', 'Dec07', 'Dec07', 'Feb08', 'Apr08', 'Jun08', 'Jul08', 'Nov08', 'Jan09',
'Mar09', 'Jul09', 'Sep09', 'Nov09', 'Dec09'],
“[‘broadband+infrastructure', 'national+security+policy']": ['Nov06'],
“[‘appropriate+investment+climate', 'e-services+products', 'broadband+infrastructure']":
['Nov09'],
“[‘national+security+policy']": ['Mar09'],
“[‘privacy+policy']": ['Apr07'],
“[‘national+security+policy', 'problem+communication', 'appropriate+investment+climate']":
['Dec09'],
“[‘e-services+products', 'appropriate+investment+climate']": ['Sep09'],
“[‘broadband+infrastructure', 'appropriate+investment+climate', 'freedom+information+policy']":
['Jun06'],
“[‘privacy+policy', 'problem+identification+mechanism', 'problem+communication']": ['Feb07'],
“[‘problem+communication']": ['Nov05', 'Nov05', 'Jul07']}

problem+communication and national+security+policy (9, 6) - {
“[‘e-services+products', 'population', 'appropriate+investment+climate']": ['Jun08'],
“[‘broadband+infrastructure', 'problem+resolution', 'e-services+products']": ['Mar09'],
“[‘broadband+infrastructure', 'population', 'e-services+products']": ['Dec06'],
“[‘problem+resolution', 'freedom+information+policy', 'population']": ['Apr08'],
“[‘broadband+infrastructure', 'e-services+products', 'population']": ['Feb06'],
“[‘problem+resolution', 'population']": ['Jul09'],
“[‘population', 'appropriate+investment+climate', 'problem+resolution']": ['Dec09'],
“[‘e-services+products', 'population', 'broadband+infrastructure']": ['Apr06'],
“[‘problem+resolution', 'broadband+infrastructure', 'privacy+policy']": ['Aug06'],
“[‘e-services+products']": ['Apr06', 'Oct06', 'Dec09'],
‘[]': ['Feb06', 'Feb06', 'Apr06', 'Apr06', 'Aug06', 'Oct06', 'Dec06', 'Aug07', 'Apr08', 'Jun08',
'Aug08', 'Mar09', 'Mar09', 'Jul09', 'Oct09', 'Dec09', 'Dec09'],
“[‘problem+resolution']": ['Oct09'],
“[‘freedom+information+policy']": ['Feb06', 'Aug08'],
“[‘problem+resolution', 'broadband+infrastructure', 'e-services+products']": ['Aug07'],
“[‘appropriate+investment+climate', 'resolution+mechanism']": ['Mar09']}

problem+resolution and broadband+infrastructure (12, 2) - {
“[‘population', 'appropriate+investment+climate', 'freedom+information+policy']": ['Jun06'],
“[‘problem+communication', 'e-services+products']": ['May06'],
“[‘appropriate+investment+climate']": ['Aug05'],
“[‘problem+communication', 'national+security+policy', 'privacy+policy']": ['Aug06'],
“[‘freedom+information+policy', 'privacy+policy']": ['Feb06'],
“[‘e-services+products']": ['Feb06'],
“[‘national+security+policy', 'population']": ['Nov06', 'Feb08'],
‘[]': ['Aug05', 'Aug05', 'Feb06', 'Feb06', 'May06', 'Jun06', 'Jun06', 'Aug06', 'Nov06', 'Jan07', 'Mar07', 'Aug07', 'Oct07', 'Feb08', 'Feb08', 'Feb08', 'Feb08', 'Apr08', 'Oct08', 'Mar09', 'Jun09', 'Nov09'],
“[‘problem+communication']": ['Apr08', 'Jun09'],
“[‘national+security+policy', 'problem+communication', 'e-services+products']": ['Aug07', 'Mar09'],
“[‘population', 'appropriate+investment+climate', 'e-services+products']": ['Nov09'],
“[‘problem+communication', 'freedom+information+policy']": ['Feb08'],
“[‘privacy+policy']": ['Oct07'],
“[‘problem+communication', 'population']": ['Mar07'],
“[‘e-services+products', 'appropriate+investment+climate']": ['Aug05'],
“[‘problem+occurence', 'appropriate+investment+climate']": ['Jan07'],
“[‘privacy+policy', 'e-services+products', 'problem+communication']": ['Oct08'],
“[‘appropriate+investment+climate', 'population', 'privacy+policy']": ['Jun06']}

resolution+mechanism and civic+stability (13, 3) - {
“[‘problem+occurence', 'privacy+policy']": ['Sep09'],
“[‘problem+identification+mechanism', 'appropriate+investment+climate', 'problem+occurence']": ['Apr07'],
“[‘national+security+policy', 'e-services+products']": ['Jun06'],
“[‘problem+occurence']": ['Dec08'],
“[‘problem+communication', 'broadband+infrastructure']": ['Oct07'],
“[‘broadband+infrastructure', 'national+security+policy', 'freedom+information+policy']": ['Dec08'],
“[‘problem+occurence', 'problem+identification+mechanism']": ['Jul07', 'Jun09'],
“[‘appropriate+investment+climate', 'problem+identification+mechanism']": ['Apr06'],
‘[]': ['Jun05', 'Apr06', 'Jun06', 'Jun06', 'Jun06', 'Oct06', 'Jan07', 'Mar07', 'Apr07', 'Apr07', 'Apr07', 'Jul07', 'Jul07', 'Aug07', 'Oct07', 'Feb08', 'Apr08', 'May08', 'Oct08', 'Dec08', 'Dec08', 'Dec08', 'Jun09', 'Jul09', 'Sep09'],
“[‘problem+identification+mechanism', 'appropriate+investment+climate']": ['Apr07'],
“[‘appropriate+investment+climate', 'problem+resolution', 'privacy+policy']": ['Jun05'],
“[‘problem+identification+mechanism', 'problem+occurence', 'freedom+information+policy']": ['Jun06'],
“[‘problem+identification+mechanism', 'problem+occurence']": ['Jul09'],
“[‘problem+identification+mechanism']": ['Jan07', 'Mar07'],
“[‘privacy+policy', 'problem+occurence', 'problem+identification+mechanism']": ['Feb08'],

“[‘freedom+information+policy', 'problem+occurence']": ['May08'],
“[‘freedom+information+policy']": ['Dec08'],
“[‘problem+occurence', 'appropriate+investment+climate']": ['Oct06']}

problem+occurence and civic+stability (11, 3) - {
“[‘problem+communication', 'e-services+products']": ['Feb08'],
“[‘problem+identification+mechanism']": ['Sep09'],
“[‘appropriate+investment+climate']": ['Jul09'],
“[‘resolution+mechanism', 'privacy+policy']": ['Sep09'],
“[‘resolution+mechanism', 'problem+identification+mechanism']": ['Jun09'],
“[‘resolution+mechanism']": ['Dec08'],
“[‘problem+identification+mechanism', 'resolution+mechanism',
'freedom+information+policy']": ['Jun06'],
“[‘problem+identification+mechanism', 'resolution+mechanism',
'appropriate+investment+climate']": ['Apr07'],
“[‘privacy+policy', 'resolution+mechanism', 'problem+identification+mechanism']": ['Feb08'],
“[‘problem+identification+mechanism', 'resolution+mechanism']": ['Jul07', 'Jul09'],
‘[]': ['Feb06', 'Feb06', 'Jun06', 'Aug06', 'Aug06', 'Oct06', 'Feb07', 'Feb07', 'Apr07', 'Jul07', 'Nov07',
'Nov07', 'Feb08', 'Feb08', 'May08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Feb09', 'Apr09', 'Jun09',
'Jun09', 'Jul09', 'Jul09', 'Sep09', 'Sep09', 'Nov09'],
“[‘privacy+policy']": ['Nov09'],
“[‘resolution+mechanism', 'appropriate+investment+climate']": ['Oct06'],
“[‘problem+resolution', 'appropriate+investment+climate', 'e-services+products']": ['Apr09'],
“[‘resolution+mechanism', 'freedom+information+policy']": ['May08'],
“[‘privacy+policy', 'freedom+information+policy']": ['Jan09']}

population and e-services+products (7, 4) - {
“[‘problem+communication', 'problem+resolution']": ['Dec07'],
“[‘broadband+infrastructure', 'national+security+policy', 'problem+communication']": ['Feb06'],
“[‘problem+resolution', 'problem+communication']": ['Dec07'],
‘[]': ['Feb06', 'Apr06', 'May06', 'Jun06', 'Dec06', 'Mar07', 'Apr07', 'Dec07', 'Dec07', 'Apr08',
'Apr08', 'Jun08', 'Jun08', 'Feb09', 'Aug09', 'Sep09', 'Nov09', 'Dec09'],
“[‘problem+resolution', 'appropriate+investment+climate']": ['Sep09'],
“[‘broadband+infrastructure', 'problem+communication', 'national+security+policy']": ['Dec06'],
“[‘national+security+policy', 'civic+stability', 'privacy+policy']": ['Aug09'],
“[‘national+security+policy']": ['Apr07'],
“[‘problem+communication', 'broadband+infrastructure', 'national+security+policy']": ['Apr06'],
“[‘national+security+policy', 'problem+resolution']": ['Jun06', 'Jun08'],
“[‘problem+communication', 'freedom+information+policy']": ['Apr08'],
“[‘broadband+infrastructure', 'problem+communication']": ['May06'],
“[‘national+security+policy', 'problem+communication', 'appropriate+investment+climate']":
['Jun08'],
“[‘problem+resolution', 'appropriate+investment+climate', 'broadband+infrastructure']":
['Nov09'],
“[‘problem+communication']": ['Dec09']}

Appendix 3 Renewable Energy term clusters

Smart Grid

hydroelectric+plant and distributed+generation (9, 8)
{"['load+management', 'smart+grid', 'demand+side+management']": ['Feb07'],
"['smart+grid', 'load+management']": ['Apr07'],
"['demand+side+management']": ['May07', 'Nov07'],
"['smart+grid', 'load+management', 'demand+side+management']": ['Jan07'],
"['smart+grid', 'demand+side+management', 'building+automation+system']": ['May06'],
'[]': ['Mar06', 'Mar06', 'May06', 'May06', 'Aug06', 'Jan07', 'Feb07', 'Apr07', 'May07', 'May07',
'Jun07', 'Jun07', 'Sep07', 'Oct07', 'Oct07', 'Nov07', 'Nov07', 'Dec07', 'Jan08', 'Apr08'],
"['building+automation+system', 'demand+side+management']": ['May06'],
"['load+management']": ['Jun07', 'Jun07', 'Sep07', 'Oct07', 'Oct07', 'Jan08', 'Apr08'],
"['smart+grid', 'demand+side+management']": ['Nov07']}

load+management and distributed+generation (10, 8)
{"['smart+grid', 'demand+side+management', 'hydroelectric+plant']": ['Feb07'],
"['hydroelectric+plant']": ['Jun07', 'Jun07', 'Sep07', 'Oct07', 'Oct07', 'Jan08', 'Apr08'],
"['smart+grid']": ['Feb07'],
'[]': ['Jan06', 'Feb06', 'Feb06', 'Jan07', 'Jan07', 'Feb07', 'Feb07', 'Feb07', 'Mar07', 'Apr07', 'Apr07',
'Jun07', 'Jun07', 'Jun07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Oct07', 'Jan08', 'Jan08', 'Feb08',
'Feb08', 'Mar08', 'Mar08', 'Apr08', 'Apr08', 'May08', 'May08', 'Jun09', 'Jun09', 'Feb10', 'Feb10'],
"['automated+meter+reading', 'Customer+Average+Interruption+Duration']": ['Jan06'],
"['hydroelectric+plant', 'smart+grid']": ['Apr07'],
"['Conservation+voltage+regulation']": ['Apr08', 'May08'],
"['smart+grid', 'hydroelectric+plant', 'demand+side+management']": ['Jan07'],
"['automated+meter+reading']": ['Feb06']}

building+automation+system and smart+grid (4, 1)
{"['distributed+generation', 'hydroelectric+plant', 'demand+side+management']": ['May06'],
"['average+system+availability']": ['Aug09', 'Sep09', 'Jan10', 'Mar10', 'May10', 'Jun10'],
"['automated+meter+reading']": ['Jan09', 'Mar09'],
'[]': ['May06', 'Jan09', 'Feb09', 'Feb09', 'Mar09', 'Jun09', 'Jun09', 'Aug09', 'Aug09', 'Sep09',
'Sep09', 'Nov09', 'Jan10', 'Jan10', 'Feb10', 'Feb10', 'Mar10', 'Mar10', 'May10', 'May10', 'Jun10',
'Jun10']}

Biofuels

E85 and ethanol (10, 8)
biofuel and bioenergy (3, 1)

E85 and cellulose (10, 4)
{"['biofuel', 'ethanol']": ['Oct07', 'May10'],
"['bioenergy', 'ethanol']": ['Jun08'],
“['bioenergy', 'ethanol', 'biomass']": ['Jul09'],

"['ethanol', 'biomass', 'bioenergy']": ['Jan08', 'Feb08', 'Mar08'],
"['ethanol', 'biofuel']": ['Dec09'],
"['biomass', 'biofuel', 'ethanol']": ['Jul08'],

'[]': ['Apr06', 'Apr06', 'May06', 'Jul06', 'Nov06', 'Nov06', 'Mar07', 'Mar07', 'Oct07', 'Dec07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Jun08', 'Jun08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Jan09', 'Mar09', 'Apr09', 'May09', 'Jul09', 'Dec09', 'Jan10', 'Feb10', 'May10'],
"['biofuel', 'biomass', 'bioenergy']": ['Nov06'],
"['biomass', 'ethanol', 'bioenergy']": ['Dec07'],
"['biofuel']": ['Nov06', 'Feb10'],
"['bioenergy']": ['Apr06', 'Mar09', 'Jan10'],
"['enzymes']": ['Apr06'],
"['biomass', 'biofuel']": ['Aug08', 'Sep08'],
"['biomass']": ['Dec07']}

E85 and ethanol (10, 8) - {"['biofuel', 'bioenergy']": ['Feb06'],
"['bioenergy', 'biofuel']": ['Aug07'],
"['bioenergy', 'biofuel', 'biomass']": ['Apr06', 'Sep08'],
"['cellulose', 'biofuel']": ['Oct07', 'Dec09'],
"['biofuel', 'bioenergy', 'biomass']": ['Jun09'],
"['solar+energy', 'bioenergy', 'biofuel']": ['Jun07'],
'[]': ['Feb06', 'Mar06', 'Mar06', 'Apr06', 'Oct06', 'Feb07', 'Apr07', 'May07', 'Jun07', 'Aug07', 'Aug07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Dec07', 'Jan08', 'Jan08', 'Feb08', 'Feb08', 'Mar08', 'Mar08', 'Apr08', 'May08', 'Jun08', 'Jul08', 'Jul08', 'Sep08', 'Nov08', 'Nov08', 'Dec08', 'Jan09', 'Jun09', 'Jul09', 'Jul09', 'Aug09', 'Sep09', 'Nov09', 'Dec09', 'Dec09', 'Jan10', 'Mar10', 'Apr10', 'May10', 'May10'],
"['biomass']": ['Oct07', 'Aug09', 'Sep09'],
"['cellulose', 'biomass', 'biofuel']": ['Jul08'],
"['biofuel', 'cellulose']": ['May10'],
"['cellulose', 'bioenergy', 'biomass']": ['Jul09'],
"['cellulose', 'bioenergy']": ['Jun08'],
"['bioenergy', 'biomass']": ['Nov08'],
"['cellulose', 'biomass', 'bioenergy']": ['Dec07', 'Jan08', 'Feb08', 'Mar08']}

ethanol and bioenergy (8, 1) - {"['solar+energy', 'biofuel', 'E85']": ['Jun07'],
"['E85', 'biofuel']": ['Aug07'],
"['biofuel', 'biomass', 'E85']": ['Jun09'],
'[]': ['Feb06', 'Apr06', 'May06', 'Jun06', 'Aug06', 'Nov06', 'Nov06', 'Jan07', 'Mar07', 'Jun07', 'Aug07', 'Nov07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Jun08', 'Jun08', 'Sep08', 'Nov08', 'Jun09', 'Jul09'],
"['cellulose', 'E85', 'biomass']": ['Dec07', 'Jan08', 'Feb08', 'Mar08'],
"['biofuel']": ['May06', 'Jun06', 'Aug06'],
"['E85', 'biomass']": ['Nov08'],
"['solar+energy', 'biofuel']": ['Jan07', 'Mar07', 'Nov07', 'Jun08'],
"['E85', 'cellulose', 'biomass']": ['Jul09'],

"['E85', 'cellulose']": ['Jun08'],
"['biofuel', 'E85', 'biomass']": ['Apr06', 'Sep08'],
"['biofuel', 'E85']": ['Feb06']}

E85 and biofuel (10, 3) - {"['cellulose', 'biomass']": ['Sep08'],
"['biomass', 'cellulose']": ['Aug08'],
"['cellulose', 'ethanol']": ['Oct07', 'Dec09', 'May10'],
"['biomass', 'bioenergy', 'cellulose']": ['Nov06'],
"['solar+energy', 'bioenergy', 'ethanol']": ['Jun07'],
"['cellulose']": ['Nov06', 'Feb10'],
'[]': ['Feb06', 'Apr06', 'Oct06', 'Nov06', 'Nov06', 'Jan07', 'Feb07', 'Apr07', 'Jun07', 'Aug07', 'Aug07', 'Oct07', 'Oct07', 'Nov07', 'Jul08', 'Aug08', 'Sep08', 'Sep08', 'Mar09', 'Jun09', 'Dec09', 'Feb10', 'May10'],
"['bioenergy', 'ethanol']": ['Feb06', 'Aug07'],
"['cellulose', 'biomass', 'ethanol']": ['Jul08'],
"['biomass', 'bioenergy']": ['Jan07'],
"['bioenergy']": ['Oct06', 'Nov07', 'Mar09'],
"['bioenergy', 'biomass', 'ethanol']": ['Apr06', 'Sep08', 'Jun09'],
"['biomass']": ['Aug07'],
"['bioenergy', 'biomass']": ['Apr07', 'Oct07']}

solar+energy and bioenergy (11, 1) - {"['ethanol', 'biofuel']": ['Jan07', 'Mar07'],
"['biofuel', 'ethanol']": ['Nov07', 'Jun08'],
"['biofuel', 'ethanol', 'E85']": ['Jun07'],
"['biofuel']": ['Oct06', 'Dec06', 'May07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Jan08', 'Feb08', 'Apr08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Mar09', 'Apr09', 'May09', 'Aug09', 'Sep09', 'Oct09', 'Nov09', 'Dec09', 'Jan10', 'Feb10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
'[]': ['Oct06', 'Dec06', 'Jan07', 'Mar07', 'May07', 'Jun07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Nov07', 'Jan08', 'Feb08', 'Apr08', 'Jun08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Mar09', 'Apr09', 'May09', 'Aug09', 'Sep09', 'Oct09', 'Nov09', 'Dec09', 'Jan10', 'Feb10', 'Mar10', 'Apr10', 'May10', 'Jun10']}

E85 and bioenergy (10, 1) - {"['biofuel', 'ethanol']": ['Feb06'],
"['solar+energy', 'biofuel', 'ethanol']": ['Jun07'],
"['biofuel', 'biomass', 'cellulose']": ['Nov06'],
"['ethanol', 'biofuel']": ['Aug07'],
"['cellulose']": ['Apr06', 'Mar09', 'Jan10'],
'[]': ['Feb06', 'Apr06', 'Apr06', 'Oct06', 'Nov06', 'Jan07', 'Jan07', 'Apr07', 'Apr07', 'May07', 'Jun07', 'Aug07', 'Oct07', 'Nov07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Jun08', 'Sep08', 'Nov08', 'Mar09', 'Mar09', 'Jun09', 'Jul09', 'Jan10', 'Feb10'],
"['cellulose', 'biomass', 'ethanol']": ['Dec07'],
"['biofuel', 'biomass']": ['Jan07', 'Apr07', 'Oct07'],
"['biofuel']": ['Oct06', 'Nov07', 'Mar09'],
"['ethanol', 'biomass']": ['Nov08'],

"['biofuel', 'biomass', 'ethanol']": ['Apr06', 'Sep08', 'Jun09'],
"['cellulose', 'ethanol']": ['Jun08'],
"['cellulose', 'ethanol', 'biomass']": ['Jan08', 'Feb08', 'Mar08', 'Jul09']}

cellulose and bioenergy (4, 1) - {"['E85', 'ethanol']": ['Jun08'],
"['E85']": ['Apr06', 'Mar09', 'Jan10'],
'[]': ['Feb06', 'Feb06', 'Mar06', 'Apr06', 'Nov06', 'Dec07', 'Dec07', 'Jan08', 'Jan08', 'Feb08', 'Feb08', 'Mar08', 'Mar08', 'Apr08', 'Apr08', 'May08', 'Jun08', 'Mar09', 'Jul09', 'Aug09', 'Aug09', 'Jan10', 'Feb10'],
"['biofuel', 'biomass']": ['Dec07'],
"['biofuel']": ['Feb06', 'Jan08', 'Feb08', 'Mar08', 'Apr08', 'May08', 'Aug09', 'Aug09', 'Feb10'],
"['E85', 'biomass', 'ethanol']": ['Dec07'],
"['E85', 'ethanol', 'biomass']": ['Jan08', 'Feb08', 'Mar08', 'Jul09'],
"['biofuel', 'E85', 'biomass']": ['Nov06']}

ethanol and biofuel (8, 3) - {"['solar+energy', 'bioenergy', 'E85']": ['Jun07'],
"['cellulose', 'E85', 'biomass']": ['Jul08'],
'[]': ['Feb06', 'Apr06', 'May06', 'Jun06', 'Jun06', 'Aug06', 'Jan07', 'Mar07', 'Jun07', 'Aug07', 'Oct07', 'Nov07', 'Jun08', 'Jul08', 'Sep08', 'Jun09', 'Dec09', 'May10'],
"['solar+energy', 'bioenergy']": ['Jan07', 'Nov07', 'Jun08'],
"['cellulose', 'E85']": ['Oct07'],
"['bioenergy']": ['May06', 'Jun06', 'Aug06'],
"['bioenergy', 'solar+energy']": ['Mar07'],
"['bioenergy', 'E85', 'biomass']": ['Apr06', 'Sep08'],
"['bioenergy', 'E85']": ['Feb06', 'Aug07'],
"['E85', 'cellulose']": ['Dec09', 'May10'],
"['bioenergy', 'biomass', 'E85']": ['Jun09']}

biofuel and bioenergy (3, 1) - {"['solar+energy', 'ethanol']": ['Jan07', 'Mar07', 'Nov07', 'Jun08'],
"['E85', 'biomass']": ['Jan07'],
"['E85', 'ethanol']": ['Feb06', 'Aug07'],
"['cellulose', 'biomass']": ['Dec07'],
"['E85']": ['Oct06', 'Nov07', 'Mar09'],
"['cellulose']": ['Feb06', 'Jan08', 'Feb08', 'Mar08', 'Apr08', 'May08', 'Aug09', 'Aug09', 'Feb10'],
'[]': ['Jan06', 'Feb06', 'Feb06', 'Mar06', 'Apr06', 'Apr06', 'May06', 'May06', 'Jun06', 'Aug06', 'Aug06', 'Sep06', 'Sep06', 'Oct06', 'Oct06', 'Nov06', 'Dec06', 'Dec06', 'Jan07', 'Jan07', 'Feb07', 'Mar07', 'Mar07', 'Apr07', 'May07', 'Jun07', 'Jun07', 'Jul07', 'Jul07', 'Aug07', 'Aug07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Nov07', 'Nov07', 'Dec07', 'Jan08', 'Jan08', 'Feb08', 'Feb08', 'Mar08', 'Apr08', 'Apr08', 'May08', 'Jun08', 'Jun08', 'Jul08', 'Jul08', 'Aug08', 'Sep08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Jan09', 'Jan09', 'Feb09', 'Mar09', 'Mar09', 'Apr09', 'Apr09', 'May09', 'May09', 'Jun09', 'Aug09', 'Aug09', 'Aug09', 'Sep09', 'Sep09', 'Oct09', 'Oct09', 'Nov09', 'Dec09', 'Dec09', 'Jan10', 'Feb10', 'Feb10', 'Mar10', 'Mar10', 'Apr10', 'Apr10', 'May10', 'Jun10', 'Jun10'],
"['biomass']": ['Sep07'],
"['solar+energy', 'ethanol', 'E85']": ['Jun07'],

"['solar+energy']": ['Oct06', 'Dec06', 'May07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Jan08', 'Feb08', 'Apr08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Mar09', 'Apr09', 'May09', 'Aug09', 'Sep09', 'Oct09', 'Nov09', 'Dec09', 'Jan10', 'Feb10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
"['biomass', 'ethanol', 'E85']": ['Jun09'],
"['enzymes']": ['Mar07'],
"['E85', 'biomass', 'cellulose']": ['Nov06'],
"['biomass', 'E85']": ['Apr07', 'Oct07'],
"['E85', 'biomass', 'ethanol']": ['Apr06', 'Sep08'],
"['ethanol']": ['May06', 'Jun06', 'Aug06']}

solar+energy and biofuel (11, 3) - {"['bioenergy', 'ethanol', 'E85']": ['Jun07'],
"['bioenergy', 'ethanol']": ['Mar07', 'Nov07', 'Jun08'],
"['bioenergy']": ['Oct06', 'Dec06', 'May07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Jan08', 'Feb08', 'Apr08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Mar09', 'Apr09', 'May09', 'Aug09', 'Sep09', 'Oct09', 'Nov09', 'Dec09', 'Jan10', 'Feb10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
'[]': ['Oct06', 'Dec06', 'Jan07', 'Mar07', 'May07', 'Jun07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Nov07', 'Jan08', 'Feb08', 'Apr08', 'Jun08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Mar09', 'Apr09', 'May09', 'Jun09', 'Jul09', 'Aug09', 'Sep09', 'Oct09', 'Nov09', 'Dec09', 'Jan10', 'Feb10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
"['ethanol', 'bioenergy']": ['Jan07']}

cellulose and biofuel (4, 3) - {"['E85', 'ethanol']": ['Oct07', 'Dec09', 'May10'],
"['E85']": ['Nov06', 'Feb10'],
'[]': ['Feb06', 'Nov06', 'Nov06', 'Oct07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Apr08', 'May08', 'Jul08', 'Aug08', 'Sep08', 'Aug09', 'Aug09', 'Dec09', 'Feb10', 'Feb10', 'May10'],
"['E85', 'biomass', 'bioenergy']": ['Nov06'],
"['E85', 'biomass', 'ethanol']": ['Jul08'],
"['bioenergy']": ['Feb06', 'Jan08', 'Feb08', 'Mar08', 'Apr08', 'May08', 'Aug09', 'Aug09', 'Feb10'],
"['E85', 'biomass']": ['Aug08', 'Sep08'],
"['bioenergy', 'biomass']": ['Dec07']}

E85 and biomass (10, 2) - {"['biofuel', 'bioenergy']": ['Jan07', 'Apr07'],
"['bioenergy', 'biofuel']": ['Oct07'],
"['biofuel', 'bioenergy', 'cellulose']": ['Nov06'],
"['cellulose', 'biofuel']": ['Aug08', 'Sep08'],
"['cellulose', 'biofuel', 'ethanol']": ['Jul08'],
"['cellulose']": ['Dec07'],
'[]': ['Apr06', 'Nov06', 'Dec06', 'Jan07', 'Apr07', 'Aug07', 'Oct07', 'Oct07', 'Dec07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Jul08', 'Aug08', 'Sep08', 'Sep08', 'Nov08', 'Jun09', 'Jul09', 'Aug09', 'Sep09'],
"['bioenergy', 'ethanol']": ['Nov08'],
"['biofuel']": ['Aug07'],
"['bioenergy', 'biofuel', 'ethanol']": ['Apr06', 'Sep08'],

"['cellulose', 'ethanol', 'bioenergy']": ['Dec07', 'Jan08', 'Feb08', 'Mar08'],
"['biofuel', 'bioenergy', 'ethanol']": ['Jun09'],
"['cellulose', 'bioenergy', 'ethanol']": ['Jul09'],
"['ethanol']": ['Oct07', 'Aug09', 'Sep09']}

Socio-economic Factors

market+efficiency and industry+efficiency (9, 4)
innovation+policy and product+process+innovation (17, 1)
labor+policy and innovation+policy (18, 17)
local+foreign+investment+policy and labor+policy (19, 18)
employment+opportunities and jobs (10, 2)

market+efficiency and industry+efficiency (9, 4) - {
“[‘innovation+policy', 'manufacture+maintenance']": ['May09'],
“[‘local+foreign+investment+policy', 'labor+policy', 'product+process+innovation']": ['Apr10'],
“[‘labor+policy', 'jobs']": ['Aug08'],
“[‘local+foreign+investment+policy']": ['Feb07', 'May07'],
“[‘jobs', 'innovation+policy', 'labor+policy']": ['Jun08'],
“[‘jobs', 'labor+policy', 'economic+activity']": ['Mar08', 'May08'],
“[‘jobs']": ['Apr06', 'Jul07', 'Oct07'],
“[‘labor+policy', 'local+foreign+investment+policy']": ['Jan08'],
“[‘economic+activity']": ['Jun07', 'Aug07', 'Sep07', 'Dec07', 'Jun08'],
“[‘employment+opportunities']": ['Jun06'],
“[‘local+foreign+investment+policy', 'market+transparency']": ['Apr10'],
“[‘local+foreign+investment+policy', 'employment+opportunities']": ['Apr07'],
“[‘jobs', 'labor+policy']": ['Jul08'], '[]': ['Apr06', 'Jun06', 'Jun06', 'Feb07', 'Apr07', 'May07',
'Jun07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Oct07', 'Dec07', 'Dec07', 'Jan08', 'Jan08', 'Mar08',
'Mar08', 'Apr08', 'Apr08', 'May08', 'May08', 'Jun08', 'Jun08', 'Jul08', 'Aug08', 'Oct08', 'Oct08',
'Jan09', 'Feb09', 'Apr09', 'May09', 'Jun09', 'Sep09', 'Oct09', 'Oct09', 'Jan10', 'Feb10', 'Feb10',
'Mar10', 'Mar10', 'Apr10', 'Apr10', 'Apr10', 'May10', 'May10', 'Jun10', 'Jun10'],
“[‘social+impact', 'employment+opportunities', 'economic+activity']": ['Jun10'],
“[‘economic+activity', 'manufacture+maintenance']": ['Sep09'],
“[‘social+service+availability', 'jobs', 'social+impact']": ['Jun09'],
“[‘income']": ['Mar08']}

visibility+communication+opportunities and production+process+automation (13, 3)
{
“[‘economic+activity', 'market+transparency', 'industry+efficiency']": ['Jun10'],
“[‘transaction+distribution+costs', 'anti-trust+industry+policy']": ['Jan08'],
“[‘transaction+distribution+costs']": ['Jan08'],
“[‘anti-trust+industry+policy', 'collaboration+knowledge+sharing']": ['Sep07', 'Dec07'],
“[‘transaction+distribution+costs', 'collaboration+knowledge+sharing']": ['Feb09'],
'[]': ['Jan07', 'Feb07', 'Mar07', 'Mar07', 'Jun07', 'Sep07', 'Nov07', 'Nov07', 'Dec07', 'Jan08', 'Jan08',
'Jan08', 'May08', 'May08', 'May08', 'Feb09', 'Apr09', 'Feb10', 'Mar10', 'Jun10'],
“[‘market+transparency', 'anti-trust+industry+policy', 'economic+activity']": ['Feb10'],

“[‘collaboration+knowledge+sharing', 'transaction+distribution+costs']": ['Nov07', 'Nov07'],
“[‘anti-trust+industry+policy', 'transaction+distribution+costs']": ['Jun07'],
“[‘economic+activity']": ['Mar10'],
“[‘market+transparency', 'innovation+policy', 'product+process+innovation']": ['Mar07'],
“[‘collaboration+knowledge+sharing']": ['Jan07', 'Feb07', 'May08', 'May08']}

innovation+policy and product+process+innovation (17, 1) - {
“[‘income', 'labor+policy', 'local+foreign+investment+policy']": ['Jun08'],
“[‘social+impact', 'employment+opportunities', 'jobs']": ['Jun09'], []': ['Feb06', 'Mar06', 'Sep06', 'Dec06', 'Feb07', 'Mar07', 'Apr07', 'May07', 'Jun07', 'Jun07', 'Aug07', 'Aug07', 'Sep07', 'Oct07', 'Oct07', 'Dec07', 'Dec07', 'Jan08', 'Feb08', 'Apr08', 'May08', 'Jun08', 'Jun08', 'Jul08', 'Sep08', 'Oct08', 'Jan09', 'Feb09', 'Apr09', 'Jun09', 'Jun09', 'Aug09', 'Sep09', 'Oct09', 'Nov09', 'Dec09', 'Dec09', 'Dec09', 'Jan10', 'Feb10', 'Mar10', 'Apr10', 'May10', 'Jun10', 'Jun10'],
“[‘production+process+automation', 'visibility+communication+opportunities', 'market+transparency']": ['Mar07'],
“[‘productivity', 'visibility+communication+opportunities']": ['Sep06'],
“[‘labor+policy', 'local+foreign+investment+policy', 'market+efficiency']": ['Mar10'],
“[‘labor+policy']": ['Aug07', 'Sep07', 'Sep09', 'Oct09'],
“[‘jobs']": ['May07'],
“[‘transaction+distribution+costs', 'local+foreign+investment+policy']": ['Apr09'],
“[‘local+foreign+investment+policy', 'labor+policy']": ['Jan10'],
“[‘market+transparency']": ['Oct08', 'Jun09'],
“[‘visibility+communication+opportunities', 'collaboration+knowledge+sharing']": ['Dec06'],
“[‘manufacture+maintenance']": ['Feb08', 'Jul08'],
“[‘local+foreign+investment+policy']": ['Dec09'],
“[‘income', 'market+efficiency']": ['Feb06'],
“[‘labor+policy', 'local+foreign+investment+policy', 'economic+activity']": ['Oct07'],
“[‘labor+policy', 'income']": ['Apr07'],
“[‘income']": ['Dec07'],
“[‘manufacture+maintenance', 'anti-trust+industry+policy', 'economic+activity']": ['Apr08'],
“[‘manufacture+maintenance', 'labor+policy', 'market+efficiency']": ['Sep08'],
“[‘income', 'labor+policy']": ['Feb07'],
“[‘labor+policy', 'local+foreign+investment+policy']": ['Oct07', 'Dec07', 'Feb09'],
“[‘income', 'transaction+distribution+costs', 'market+transparency']": ['Dec09'],
“[‘manufacture+maintenance', 'economic+activity']": ['May08']}

market+efficiency and jobs (9, 2) - {
“[‘employment+opportunities', 'local+foreign+investment+policy']": ['Feb07'], []': ['Apr06', 'May06', 'Oct06', 'Dec06', 'Feb07', 'Feb07', 'Apr07', 'Jul07', 'Jul07', 'Oct07', 'Mar08', 'May08', 'Jun08', 'Jul08', 'Aug08', 'Jun09', 'Jun09', 'Nov09'],
“[‘industry+efficiency']": ['Apr06', 'Jul07', 'Oct07'],
“[‘employment+opportunities']": ['May06', 'Oct06', 'Dec06', 'Feb07', 'Apr07'],
“[‘industry+efficiency', 'labor+policy', 'economic+activity']": ['Mar08', 'May08'],
“[‘industry+efficiency', 'innovation+policy', 'labor+policy']": ['Jun08'],
“[‘industry+efficiency', 'labor+policy']": ['Jul08', 'Aug08'],

“[‘social+service+availability', 'social+impact', 'industry+efficiency']": ['Jun09'],
“[‘social+service+availability', 'transaction+distribution+costs']": ['Jun09']}

labor+policy and product+process+innovation (18, 1) - {
“[‘innovation+policy', 'income']": ['Feb07', 'Apr07'],
“[‘innovation+policy', 'local+foreign+investment+policy', 'economic+activity']": ['Oct07'],
“[‘local+foreign+investment+policy']": ['Dec08', 'Apr09'],
“[‘innovation+policy']": ['Aug07', 'Sep07', 'Sep09', 'Oct09'],
“[‘income', 'innovation+policy', 'local+foreign+investment+policy']": ['Jun08'],
“[‘local+foreign+investment+policy', 'social+impact', 'employment+opportunities']": ['Nov08'],
'[]': ['Feb07', 'Apr07', 'Aug07', 'Aug07', 'Sep07', 'Oct07', 'Oct07', 'Dec07', 'Dec07', 'Jun08',
'Sep08', 'Nov08', 'Dec08', 'Feb09', 'Apr09', 'Apr09', 'Sep09', 'Oct09', 'Jan10', 'Mar10', 'Apr10'],
“[‘market+efficiency', 'local+foreign+investment+policy', 'industry+efficiency']": ['Apr10'],
“[‘income', 'economic+activity', 'local+foreign+investment+policy']": ['Aug07'],
“[‘manufacture+maintenance', 'market+efficiency', 'innovation+policy']": ['Sep08'],
“[‘innovation+policy', 'local+foreign+investment+policy', 'market+efficiency']": ['Mar10'],
“[‘innovation+policy', 'local+foreign+investment+policy']": ['Oct07', 'Dec07', 'Feb09', 'Jan10'],
“[‘economic+activity', 'income']": ['Dec07']}

anti-trust+industry+policy and market+transparency (20, 12) - {
“[‘manufacture+maintenance']": ['Apr07'],
“[‘visibility+communication+opportunities']": ['Jun07'],
“[‘transaction+distribution+costs']": ['Jan08'],
“[‘productivity']": ['Jun08', 'Aug08', 'Sep08', 'Jan10'],
“[‘production+process+automation']": ['Jun07'], '[]': ['Oct06', 'Nov06', 'Apr07', 'Apr07', 'May07',
'May07', 'Jun07', 'Jun07', 'Jul07', 'Nov07', 'Nov07', 'Dec07', 'Jan08', 'Jun08', 'Jun08', 'Aug08',
'Sep08', 'Dec08', 'May09', 'Dec09', 'Jan10', 'Feb10', 'Mar10', 'Apr10'],
“[‘visibility+communication+opportunities', 'income', 'innovation+policy']": ['Oct06'],
“[‘visibility+communication+opportunities', 'economic+activity',
'production+process+automation']": ['Feb10'],
“[‘labor+policy', 'production+process+automation', 'local+foreign+investment+policy']":
['Nov06'],
“[‘labor+policy', 'local+foreign+investment+policy', 'innovation+policy']": ['Jul07']}

labor+policy and innovation+policy (18, 17) - {
“[‘income', 'product+process+innovation']": ['Feb07', 'Apr07'],
“[‘anti-trust+industry+policy']": ['Mar07'],
“[‘income', 'local+foreign+investment+policy', 'product+process+innovation']": ['Jun08'],
“[‘product+process+innovation']": ['Aug07', 'Sep07', 'Sep09', 'Oct09'], '[]': ['Feb07', 'Feb07',
'Mar07', 'Mar07', 'Apr07', 'Apr07', 'May07', 'Jun07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Oct07',
'Dec07', 'Feb08', 'May08', 'Jun08', 'Jun08', 'Sep08', 'Feb09', 'Sep09', 'Sep09', 'Oct09', 'Jan10',
'Mar10'],
“[‘economic+activity', 'market+transparency']": ['Sep09'],
“[‘product+process+innovation', 'local+foreign+investment+policy', 'market+efficiency']":
['Mar10'],

“[‘manufacture+maintenance', 'product+process+innovation', 'market+efficiency']": ['Sep08'],
“[‘product+process+innovation', 'local+foreign+investment+policy']": ['Oct07', 'Dec07', 'Feb09', 'Jan10'],
“[‘local+foreign+investment+policy', 'market+transparency']": ['Mar07'],
“[‘economic+activity', 'local+foreign+investment+policy']": ['Feb08', 'May08'],
“[‘local+foreign+investment+policy', 'market+transparency', 'anti-trust+industry+policy']": ['Jul07'],
“[‘jobs', 'industry+efficiency', 'market+efficiency']": ['Jun08'],
“[‘income']": ['Feb07', 'May07', 'Jun07'],
“[‘product+process+innovation', 'local+foreign+investment+policy', 'economic+activity']": ['Oct07']}

transaction+distribution+costs and collaboration+knowledge+sharing (6, 5) - {
“[‘visibility+communication+opportunities']": ['May07', 'Jul07', 'Jul08', 'Dec08'],
“[‘anti-trust+industry+policy', 'innovation+policy', 'production+process+automation']": ['Mar06'],
“[‘innovation+policy']": ['Jul07'],
“[‘production+process+automation']": ['Jan06', 'Sep07', 'Feb08'],
“[‘visibility+communication+opportunities', 'production+process+automation']": ['Nov07', 'Nov07'], '[]': ['Jan06', 'Mar06', 'Dec06', 'May07', 'Jul07', 'Jul07', 'Sep07', 'Nov07', 'Nov07', 'Feb08', 'Feb08', 'Mar08', 'Jul08', 'Dec08', 'Dec08', 'Feb09', 'Jan10', 'Feb10', 'Feb10', 'Apr10', 'Apr10'],
“[‘production+process+automation', 'visibility+communication+opportunities']": ['Feb09'],
“[‘anti-trust+industry+policy', 'visibility+communication+opportunities']": ['Mar08']}

collaboration+knowledge+sharing and production+process+automation (5, 3) - {
“[‘transaction+distribution+costs', 'anti-trust+industry+policy', 'innovation+policy']": ['Mar06'],
“[‘visibility+communication+opportunities']": ['Jan07', 'Feb07', 'May08', 'May08'],
“[‘visibility+communication+opportunities', 'anti-trust+industry+policy']": ['Dec07'],
“[‘visibility+communication+opportunities', 'transaction+distribution+costs']": ['Nov07', 'Nov07'],
“[‘transaction+distribution+costs']": ['Jan06', 'Sep07', 'Feb08'],
“[‘anti-trust+industry+policy', 'visibility+communication+opportunities']": ['Sep07'], '[]': ['Jan06', 'Jan06', 'Mar06', 'May06', 'Jan07', 'Jan07', 'Feb07', 'Jun07', 'Sep07', 'Sep07', 'Nov07', 'Nov07', 'Nov07', 'Dec07', 'Feb08', 'May08', 'May08', 'Jun08', 'Oct08', 'Oct08', 'Feb09', 'Dec09'],
“[‘transaction+distribution+costs', 'visibility+communication+opportunities']": ['Feb09']}

social+impact and social+service+availability (11, 8) - {
“[‘market+efficiency', 'jobs', 'industry+efficiency']": ['Jun09'],
“[‘jobs']": ['Jan09'],
“[‘employment+opportunities']": ['Feb09'], '[]': ['Feb08', 'Mar08', 'Apr08', 'May08', 'Jun08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Jun09', 'Sep09', 'Jan10', 'Mar10', 'Apr10', 'May10']}

visibility+communication+opportunities and collaboration+knowledge+sharing (13, 5) - {

“[‘production+process+automation', 'transaction+distribution+costs']": ['Nov07', 'Feb09'],
“[‘anti-trust+industry+policy']": ['Aug07'],
“[‘anti-trust+industry+policy', 'production+process+automation']": ['Sep07', 'Dec07'],
“[‘transaction+distribution+costs', 'production+process+automation']": ['Nov07'],
“[‘production+process+automation']": ['Jan07', 'Feb07', 'May08', 'May08'],
“[‘transaction+distribution+costs']": ['May07', 'Jul07', 'Jul08', 'Dec08'],
“[‘anti-trust+industry+policy', 'transaction+distribution+costs']": ['Mar08'],
“[‘market+transparency']": ['Jun07'],
'[]': ['Nov06', 'Dec06', 'Jan07', 'Feb07', 'Feb07', 'May07', 'May07', 'Jun07', 'Jul07', 'Aug07', 'Aug07', 'Sep07', 'Oct07', 'Oct07', 'Nov07', 'Nov07', 'Dec07', 'Mar08', 'May08', 'May08', 'Jul08', 'Aug08', 'Aug08', 'Nov08', 'Dec08', 'Feb09'],

“[‘innovation+policy', 'product+process+innovation']": ['Dec06']}
local+foreign+investment+policy and labor+policy (19, 18) - {
“[‘production+process+automation']": ['Apr06'], '[]': ['Jan06', 'Apr06', 'May06', 'Jun06', 'Aug06', 'Aug06', 'Sep06', 'Oct06', 'Nov06', 'Dec06', 'Mar07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Oct07', 'Oct07', 'Oct07', 'Dec07', 'Dec07', 'Jan08', 'Feb08', 'May08', 'Jun08', 'Jun08', 'Nov08', 'Dec08', 'Jan09', 'Feb09', 'Feb09', 'Apr09', 'Oct09', 'Nov09', 'Jan10', 'Feb10', 'Mar10', 'Apr10'],
“[‘market+transparency', 'innovation+policy']": ['Mar07'],
“[‘product+process+innovation', 'innovation+policy']": ['Oct07', 'Dec07', 'Feb09', 'Jan10'],
“[‘market+efficiency', 'industry+efficiency']": ['Jan08'],
“[‘product+process+innovation']": ['Dec08', 'Apr09'],
“[‘social+impact', 'product+process+innovation', 'employment+opportunities']": ['Nov08'],
“[‘market+efficiency', 'product+process+innovation', 'industry+efficiency']": ['Apr10'],
“[‘market+transparency']": ['May06', 'Jun06', 'Sep06'],
“[‘income', 'economic+activity', 'product+process+innovation']": ['Aug07'],
“[‘productivity', 'market+transparency']": ['Oct09'],
“[‘manufacture+maintenance', 'income']": ['Aug06'],
“[‘economic+activity', 'income']": ['Nov09'],
“[‘economic+activity']": ['Feb09'],
“[‘employment+opportunities']": ['Sep07'],
“[‘market+transparency', 'anti-trust+industry+policy', 'innovation+policy']": ['Jul07'],
“[‘employment+opportunities', 'social+impact', 'industry+efficiency']": ['Oct07'],
“[‘anti-trust+industry+policy']": ['Jan09'],
“[‘production+process+automation', 'anti-trust+industry+policy', 'market+transparency']": ['Nov06'],
“[‘product+process+innovation', 'innovation+policy', 'market+efficiency']": ['Mar10'],
“[‘income', 'manufacture+maintenance']": ['Jan06'],
“[‘economic+activity', 'innovation+policy']": ['Feb08', 'May08'],
“[‘income', 'innovation+policy', 'product+process+innovation']": ['Jun08'],
“[‘product+process+innovation', 'innovation+policy', 'economic+activity']": ['Oct07']}

employment+opportunities and jobs (10, 2) - {
“[‘local+foreign+investment+policy']": ['Apr07'],
“[‘innovation+policy', 'social+impact', 'product+process+innovation']": ['Jun09'],

“[‘market+efficiency']": ['May06', 'Oct06', 'Dec06', 'Feb07', 'Apr07'],
“[‘labor+policy']": ['Apr10'],
“[‘social+impact', 'innovation+policy']": ['Feb08'],
“[‘social+impact']": ['May09'], '[]': ['May06', 'Aug06', 'Sep06', 'Oct06', 'Oct06', 'Dec06', 'Dec06', 'Feb07', 'Feb07', 'Apr07', 'Apr07', 'Aug07', 'Sep07', 'Nov07', 'Dec07', 'Jan08', 'Feb08', 'Apr08', 'May08', 'Jun08', 'Jul08', 'Aug08', 'Sep08', 'Dec08', 'Jan09', 'Mar09', 'Apr09', 'May09', 'Jun09', 'Jul09', 'Aug09', 'Oct09', 'Nov09', 'Jan10', 'Feb10', 'Mar10', 'Apr10', 'Apr10', 'May10', 'Jun10'],
“[‘innovation+policy', 'social+impact']": ['Apr08'],
“[‘local+foreign+investment+policy', 'market+efficiency']": ['Feb07']}

Energy Consumption

security and energy+service+provider (7, 3)
service+area and security (8, 7)

security and energy+service+provider (7, 3) - {
“[‘service+area']": ['Feb07', 'Apr07', 'May07', 'Jun07', 'Aug07', 'Sep07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Apr08', 'May08', 'Jul08', 'Aug08', 'Sep08', 'Nov08', 'Dec08', 'Jan09', 'Apr09', 'May09', 'Sep09', 'Nov09', 'Dec09', 'Jan10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
“[‘residential']": ['Jun09'], '[]': ['Feb07', 'Apr07', 'May07', 'May07', 'Jun07', 'Jun07', 'Aug07', 'Aug07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Dec07', 'Dec07', 'Jan08', 'Jan08', 'Feb08', 'Mar08', 'Mar08', 'Apr08', 'May08', 'May08', 'Jul08', 'Aug08', 'Sep08', 'Nov08', 'Dec08', 'Jan09', 'Apr09', 'May09', 'May09', 'Jun09', 'Jun09', 'Jun09', 'Sep09', 'Oct09', 'Nov09', 'Nov09', 'Dec09', 'Jan10', 'Mar10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
“[‘service+area', 'residential']": ['Jun09'],
“[‘service+area', 'asset+management']": ['May07', 'Jun07', 'Aug07', 'Sep07', 'Oct07', 'Dec07'],
“[‘residential', 'asset+management', 'service+area']": ['Jun09'],
“[‘residential', 'asset+management']": ['Oct09']}

security and residential (7, 6) - {
“[‘reliability', 'service+area']": ['Jul07'],
“[‘service+area']": ['Apr06', 'Jan07', 'May07', 'Jun07', 'Aug07', 'Sep07', 'Jan08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Mar09', 'Jun10'],
“[‘service+area', 'asset+management', 'reliability']": ['Oct07'],
“[‘energy+service+provider', 'service+area']": ['Jun09'],
“[‘energy+service+provider']": ['Jun09'],
“[‘rate+of+return', 'service+area']": ['Feb08', 'Mar08'], '[]': ['Apr06', 'Jan07', 'May07', 'Jun07', 'Jul07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Mar09', 'May09', 'Jun09', 'Jun09', 'Jun09', 'Oct09', 'Jun10'],
“[‘energy+service+provider', 'asset+management']": ['Oct09'],
“[‘service+area', 'reliability', 'asset+management']": ['May09'],
“[‘reliability', 'service+area', 'asset+management']": ['Dec07'],
“[‘service+area', 'reliability']": ['Jul07'],
“[‘energy+service+provider', 'asset+management', 'service+area']": ['Jun09'],
“[‘service+area', 'asset+management']": ['Jul08', 'Aug08']}

service+area and residential (8, 6) - {
“[‘security', 'asset+management', 'reliability']": ['Oct07'],
“[‘energy+service+provider']": ['Oct09'],
“[‘security', 'reliability']": ['Jul07', 'Jul07'], '[]': ['Apr06', 'Jan07', 'May07', 'Jun07', 'Jul07', 'Jul07', 'Aug07', 'Sep07', 'Oct07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Jul08', 'Aug08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Mar09', 'May09', 'Jun09', 'Jun09', 'Oct09', 'Oct09', 'Jun10'],
“[‘security', 'energy+service+provider']": ['Jun09'],
“[‘security']": ['Apr06', 'Jan07', 'May07', 'Jun07', 'Aug07', 'Sep07', 'Jan08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Mar09', 'Jun10'],
“[‘asset+management', 'rate+of+return']": ['Oct09'],
“[‘security', 'asset+management']": ['Jul08', 'Aug08'],
“[‘energy+service+provider', 'asset+management', 'security']": ['Jun09'],
“[‘rate+of+return', 'security']": ['Feb08', 'Mar08'],
“[‘reliability', 'security', 'asset+management']": ['Dec07', 'May09']}

service+area and security (8, 7) - {
“[‘residential', 'reliability']": ['Jul07', 'Jul07'],
“[‘residential', 'asset+management', 'reliability']": ['Oct07'],
“[‘energy+service+provider']": ['Feb07', 'Apr07', 'May07', 'Jun07', 'Aug07', 'Sep07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Apr08', 'May08', 'Jul08', 'Aug08', 'Sep08', 'Nov08', 'Dec08', 'Jan09', 'Apr09', 'May09', 'Sep09', 'Nov09', 'Dec09', 'Jan10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
“[‘residential']": ['Apr06', 'Jan07', 'May07', 'Jun07', 'Aug07', 'Sep07', 'Jan08', 'Sep08', 'Oct08', 'Nov08', 'Dec08', 'Mar09', 'Jun10'],
“[‘energy+service+provider', 'residential', 'asset+management']": ['Jun09'],
“[‘residential', 'rate+of+return']": ['Feb08', 'Mar08'], '[]': ['Jan06', 'Jan06', 'Feb06', 'Feb06', 'Mar06', 'Mar06', 'Apr06', 'Apr06', 'May06', 'May06', 'Jun06', 'Jun06', 'Aug06', 'Aug06', 'Oct06', 'Oct06', 'Dec06', 'Dec06', 'Jan07', 'Jan07', 'Feb07', 'Feb07', 'Mar07', 'Mar07', 'Apr07', 'Apr07', 'May07', 'May07', 'May07', 'Jun07', 'Jun07', 'Jun07', 'Jul07', 'Jul07', 'Jul07', 'Aug07', 'Aug07', 'Aug07', 'Sep07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Nov07', 'Nov07', 'Dec07', 'Dec07', 'Dec07', 'Jan08', 'Jan08', 'Feb08', 'Feb08', 'Feb08', 'Mar08', 'Mar08', 'Mar08', 'Apr08', 'Apr08', 'May08', 'May08', 'Jun08', 'Jun08', 'Jul08', 'Jul08', 'Aug08', 'Aug08', 'Sep08', 'Sep08', 'Oct08', 'Oct08', 'Nov08', 'Nov08', 'Dec08', 'Dec08', 'Jan09', 'Jan09', 'Feb09', 'Mar09', 'Mar09', 'Apr09', 'May09', 'May09', 'Jun09', 'Jun09', 'Jul09', 'Jul09', 'Aug09', 'Aug09', 'Sep09', 'Nov09', 'Dec09', 'Jan10', 'Feb10', 'Feb10', 'Mar10', 'Mar10', 'Apr10', 'Apr10', 'May10', 'May10', 'Jun10', 'Jun10'],
“[‘rate+of+return']": ['Apr07'],
“[‘reliability', 'residential', 'asset+management']": ['Dec07', 'May09'],
“[‘energy+service+provider', 'residential']": ['Jun09'],
“[‘energy+service+provider', 'asset+management']": ['May07', 'Jun07', 'Aug07', 'Sep07', 'Oct07', 'Dec07'],
“[‘residential', 'asset+management']": ['Jul08', 'Aug08']}

service+area and energy+service+provider (8, 3) - {
“[‘residential']": ['Oct09'],
“[‘security', 'residential']": ['Jun09'], '[]': ['Feb07', 'Apr07', 'May07', 'May07', 'Jun07', 'Jun07', 'Aug07', 'Aug07', 'Sep07', 'Sep07', 'Oct07', 'Dec07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Apr08',

'May08', 'Jul08', 'Aug08', 'Sep08', 'Nov08', 'Dec08', 'Jan09', 'Apr09', 'May09', 'Jun09', 'Jun09', 'Sep09', 'Oct09', 'Nov09', 'Dec09', 'Jan10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
“[‘security']": ['Feb07', 'Apr07', 'May07', 'Jun07', 'Aug07', 'Sep07', 'Dec07', 'Jan08', 'Feb08', 'Mar08', 'Apr08', 'May08', 'Jul08', 'Aug08', 'Sep08', 'Nov08', 'Dec08', 'Jan09', 'Apr09', 'May09', 'Sep09', 'Nov09', 'Dec09', 'Jan10', 'Mar10', 'Apr10', 'May10', 'Jun10'],
“[‘security', 'asset+management']": ['May07', 'Jun07', 'Aug07', 'Sep07', 'Oct07', 'Dec07'],
“[‘residential', 'asset+management', 'security']": ['Jun09']}

residential and asset+management (6, 1) - {
“[‘reliability']": ['Oct06'],
“[‘rate+of+return', 'service+area']": ['Oct09'],
“[‘energy+service+provider', 'service+area', 'security']": ['Jun09'],
“[‘service+area', 'security']": ['Jul08', 'Aug08'],
“[‘reliability', 'service+area', 'security']": ['Dec07'],
“[‘energy+service+provider', 'security']": ['Oct09'],
“[‘service+area', 'reliability', 'security']": ['May09'], []': ['Jan06', 'Jan06', 'Feb06', 'Feb06', 'Mar06', 'Jun06', 'Jun06', 'Aug06', 'Sep06', 'Sep06', 'Oct06', 'Oct06', 'Dec06', 'Dec06', 'Jan07', 'Mar07', 'Oct07', 'Dec07', 'Jul08', 'Aug08', 'May09', 'Jun09', 'Oct09', 'Oct09'],
“[‘rate+of+return']": ['Feb06', 'Dec06'],
“[‘service+area', 'security', 'reliability']": ['Oct07'],
“[‘deregulation']": ['Feb06', 'Mar06', 'Jun06', 'Aug06', 'Oct06']}

References

Anjewierden, A. and Efimova, L. Understanding Weblog Communities Through Digital Traces: A Framework, a Tool and an Example." In Proc. of the OTM 2006 Workshops, LNCS 4277, pp 279–289. Springer, 2006.

Berry, M. W., Dumais, S.T., and Letsche, T. A. (1995). Computational Methods for Intelligent Information Access. In Proceedings of the 1995 ACM/IEEE Supercomputing Conference.

Bjorneborn, L., and Ingwersen, P. (2004) "Toward a basic framework for webometrics", Journal of the American Society for Information Science and Technology. Vol. 55, No. 14, pages 1216–1227.

Blei, D. M., Ng, A., and Jordan, M. "Latent dirichlet allocation." J. Mach. Learn. Res. Vol. 3. March 2003, 993-1022.

Boyack, K. (2002) "Domain visualization using VxInsight® for science and technology management", Journal of the American Society for Information Science and Technology, Vol. 53, No. 9, 764-774.

Breslin, J.G. et. Al. Towards Semantically-Interlinked Online Communities." The Semantic Web: Research and Applications. Springer, 2005, pp 500-514.

Cachia, R. et. al. (2007) "Grasping the potential of online social networks for foresight." Technological Forecasting and Social Change. Vol. 74, pp 1179-1203.

Chakraborti, S. Mukras, R., Lothian, R., Wiratunga, N., Watt, S., and Harper, D.. (2007) "Supervised Latent Semantic Indexing Using Adaptive Sprinkling". Proceedings of IJCAI, pages 1582–1587.

Chen, Y. Tsai, F.S., and Chan, K.L. Blog Search and Mining in the Business Domain." In the 2007 ACM SIGKDD Workshop on Domain Driven Data Mining. ACM, 2007.

Chiesa, M. Masdar Institute of Science and Technology. Personal Interview, July 2010.

Crossley, S., T. Salsbury, P. McCarthy & D. McNamara （2008）. "Using Latent Semantic Analysis to explore second language lexical development". Proceedings of the Twenty-first International FLAIRS Conference.

Courseault, C. (2004) "A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions". PhD Thesis, Alan Porter (Advisor), Georgia Institute of Technology.

Daim, T. et. al. (2006) "Forecasting emerging technologies: Use of bibliometrics and patent analysis", Technological Forecasting & Social Change, Vol. 73, 981–1012.

Daim, T.U., Rueda, G.R., Martin, H.T. (2005) "Technology forecasting using bibliometric analysis and system dynamics", Technology Management: A Unifying Discipline for Melting the Boundaries, Jul. 2005.

Dasgupta, A.,Kumar, Raghavan, and Tomkins. (2005) Variable latent semantic indexing. In KDD ’05, pp. 13–21. ACM Press.

de Miranda Santo, M., et. al. (2006) "Text mining as a valuable tool in foresight exercises: A study on nanotechnology", Technological Forecasting and Social Change, Vol. 73, No. 8, Pages 1013-1027.

Emziane, M. Masdar Institute of Science and Technology. Personal Interview, July 2010.

Farahat, A., and Chen, F. (2006) “Improving Probabilistic Latent Semantic Analysis with Principal Component Analysis”, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy.

Glance, N.S., Hurst, M. Tomokiyo, T. BlogPulse: Automated Trend Discovery for Weblogs,” In WWW

Glanzel, W. (1996) "The need for standards in bibliometric research and technology", Scientometrics, Vol. 35, No. 2.

Gorman, J.C., Foltz, P.W., Kiekel, P.A., Martin, M. J., & Cooke, N. J. (2003). "Evaluation of latent semantic analysis-based measures of team communications". Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting, 424-428.

Hashaikeh, R. Masdar Institute of Science and Technology. Personal Interview, July 2010.

Hauptman, O. and Pope, S. (1979) "The process of applied technology forecasting: A study of executive analysis, anticipation, and planning", Technological Forecasting and Social Change, Vol. 42, No. 2, Sep. 1992, 193-211.

Hjorland, B., Albrechtsen, H. (1995) "Toward a New Horizon in Information Science: Domain-Analysis", Journal of the American Society for Information Science, Vol. 46, No. 6.

Hummon, N., Dereian, P., (1989) "Connectivity in a citation network: The development of DNA theory", Social Networks, Elsevier.

Kalz, M., Van Bruggen, J., Giesbers, B., & Koper, R. (2007). Prior Learning Assessment with Latent Semantic Analysis. In F. Wild, M. Kalz, J. Van Bruggen & R. Koper (Eds.). Proceedings of the First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning (pp. 24-25). Heerlen, The Netherlands: Open University of the Netherlands.

Kaye, B.K. It’s a blog, blog, blog world: Users and users of weblogs.” Atlantic Journal of Communication, 13(2), pp 73-95. 2005

Koh, H. and Magee, C. (2006), "A functional approach for studying technological progress: Application to information technology", Technological Forecasting and Social Change, Vol. 73, No. 9, Nov. 2006, 1061-1083.

Kostoff, R. et. al. (2000) "Fullerene Data Mining Using Bibliometrics and Database Tomography", Journal of Chemical Information and Computer Sciences, Vol. 40, No. 1, 19-39.

Kullback, S. Information Theory and Statistics, Dover Publications Inc., Mineola, New York, 1968.

Kumar, R., et.al. On the Bursty Evolution of Blogspace." World Wide Web. 8(2), 2005.

Lancaster, F., Lee, J., (1985) "Bibliometric Techniques Applied to Issues Management: A Case Study", Journal of the American Society for Information Science, Vol. 36, No. 6.

Lenhart, A. and Fox, S. Bloggers." Pew Internet & American Life Project, Jul. 2006

Macdonald, C. and Ounis, I. (2006) "The TREC Blog06 Collection: Creating and Analysing a Blog Test Collection", DCS Technical Report TR-2006-224. Univ. of Glasgow.

Makridakis, S. and Hibon, M. (1979), "Accuracy of Forecasting: An Empirical Investigation", J. R. Statistical Soc., Part 2, Vol. 142, 1979, 97-145.

Mann, G., Mimno, D., McCallum, A. (2006) "Bibliometric impact measures leveraging topic analysis", JCDL '06 Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries.

Mishne, G., Glance, N. Leave a Reply: An Analysis of Weblog Comments." ," In WWW

Mosier, C. "A Critical Examination of the Concepts of Face Validity". Educational and Psychological Measurement, 7(2), July 1947.

Narin, F. (1996). "Bibliometric Performance Measures", Scientometrics, Vol. 36, No. 3. 397-420.

Nerur, S. et. al. (2005) "Assessing the relative influence of journals in a citation network", Communications of the ACM. Vol. 48, No. 11.

Nielsen (2006). Nielsen Buzzmetrics blog dataset, International Conference on Weblogs and Social Media, Boulder CO.

Okubo, Y. (1997), "Bibliometric Indicators and Analysis of Research Systems: Methods and Examples", OECD Science, Technology and Industry Working Papers, 1997/1, OECD Publishing.

Osorio-Urzua, C. Architectural Innovation, Functional Emergence and Diversification in Engineering Systems." Ph.D. Thesis, MIT Engineering Systems Division, 2007.

Prieto, R. Masdar Institute of Science and Technology. Personal Interview, July 2010.

Pritchard, A. (1969), "Statistical bibliography or bibliometrics?", Journal of Documentation, Vol. 25, No. 4, 348-349.

Roberts, E. (1969), "Exploratory and normative technological forecasting: A critical appraisal", Technological Forecasting, Vol. 1, No. 2, Autumn 1969, 113-127.

Sgouridis, S. Masdar Institute of Science and Technology. Personal Interview, July 2010.

Small, H. (1973) "Co-citation in the Scientific Literature", Journal of the American Society of Information Science, Jul-Aug 1973.

Sussman, et. al. "The CLIOS Process: A User's Guide." MIT Teaching Note, April 26, 2007.

Tabaei, A. Masdar Institute of Science and Technology. Personal Interview, July 2010.

Thomson Scientific Announces Strategic Alliance with Collexis to Develop Custom Data Mining Solution for Web of Science Users," Thompson Reuters press release, Feb. 21, 2008

Valerdi, R., Ross, A., Rhodes, D. A Framework for Evolving System of Systems Engineering." CrossTalk. Oct. 2007.

Van Raan, A.F.J. (1996), "Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises", Scientometrics, Vol. 36, No. 3. 397-420.

Vantage Point, http://www.thevantagepoint.com

Verbeek, A. et. al. (2002). "Measuring progress and evolution in science and technology - I: The multiple uses of bibliometric indicators", International Journal of Management Reviews, Vol. 4, No. 2. 179-211.

Yoon, B. and Park, Y., (2005) "A systematic approach for identifying technology opportunities: Keyword-based morphology analysis", Technological Forecasting and Social Change, Vol. 72, No. 2, Feb. 2005, Pages 145-160.

Zeineldin, H. Masdar Institute of Science and Technology. Personal Interview, July 2010.

Ziegler, Blaine, Woon, Wei Lee and Madnick, Stuart E., Latent Semantic Analysis Applied to Tech Mining. MIT Sloan Research Paper No. 4726-09. Sept. 2009. Available at SSRN: http://ssrn.com/abstract=1356011

Zukas, A., R. J. Price, (2003) "Document Categorization Using Latent Semantic Indexing", in: Proceedings, Symposium on Document Image Understanding Technology, pp. 87-91.

2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004, at WWW'04: the 13th international conference on World Wide Web, 2004.

2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at WWW'06: the 15th international conference on World Wide Web, 2006.