Emergence of breakthroughs: a case study of RNA interference

Sen Chai
Wyss House
Harvard Business School
Soldiers Field Rd., Boston MA 02163
schai@hbs.edu
650 235 6178

Lee Fleming
Morgan 485
Harvard Business School
Soldiers Field Rd., Boston MA 02163
lfleming@hbs.edu
617 495 6613

January 29, 2011

We would like to thank the Harvard Business School Department of Research and the National Science Foundation, Grant #0965259, for supporting our work. Errors and omissions remain ours.

Introduction

Breakthrough or radical inventions have shown to be an important source of technological advancement, wealth creation and economic growth. They are sources of creative destruction that bring about paradigm shifts, new technological trajectories, and require different competencies. They are at the core of wealth creation as evidenced by the empirical finding that the top 10% of patents collect 48 to 93% of financial payoffs (Scherer & Harhoff, 2000). The economic impact of breakthroughs is at the crux of research aimed to establish sources of radical inventions. At the organizational level, starting with Schumpeter's self-conflicting arguments with regard to the optimal organizational structure, scholars have long debated whether entrants (Schumpeter, 1942) or incumbents (Ahuja & Lampert, 2001) are more likely sources of breakthrough inventions. At more micro level of analyses, scholars have studied whether lone individuals versus teams are most at risk of inventing breakthroughs (Singh & Fleming, 2010).

The present paper contributes to this line of research by investigating two research questions: where scientific breakthroughs arise from within a community and what characteristics make a particular scientist more likely to discover a breakthrough. Three major aspects differentiate this project from previous papers in the literature: first, it focuses on scientific breakthroughs as opposed to technological breakthroughs; second, it shifts from the organizational unit of analysis to the individual; third, it brings data to a question which has previously focused on the publication or patent as the unit of analysis, or remained mainly theoretical; and fourth, it attempts to use data in the years prior to breakthrough to predict the breakthrough.

The focus on scientific breakthroughs reflects the increasing role science plays in spurring technological advancement and economic growth. Indeed, several empirical studies

have confirmed the link between science and economic growth where cumulative published research accelerates growth (Adams, 1990), as well as the link between science and technological innovation where increased university research spending is associated with greater rates of local patenting (Jaffe & Trajtenberg, 1996). Research merging these ideas has demonstrated that science serves as a guide in the search for technological progress; it leads inventors more directly to promising combinations thus revealing a precise mechanism through which science accelerates the rate of technological innovation (Fleming & Sorenson, 2004).

Given science's crucial role in advancing economic growth, understanding the process of scientific knowledge creation is vital especially from both a managerial and policy standpoint. This research will aid science-based firms in producing more at risk breakthrough research, and generally help firms and their managers identify commercializable opportunities. From a policy standpoint, this project identifies scientists at risk of breakthrough, which is a first step in eventually moving up levels of analysis to locating communities of scientists more likely to discover breakthroughs and, thus, enabling more targeted governmental subsidies and private investments into them. Furthermore, aside from the purely economic motivations for studying breakthroughs, they are also sources of social benefit that enhance social welfare (Trajtenberg, 1990).

The current paper enables us to gather empirical evidence quantitatively by exploring our research questions through econometric analysis of a large dataset – the disambiguated Pubmed Author-ity database (Torvik & Smalheiser, 2009). Furthermore, we plot network maps as an excellent tool to graphically and structurally visualize collaborative networks prior to the emergence of breakthrough.

We employ two outcome variables that measure productivity and impact of academic publications. They consist, respectively, of the number of publications during the breakthrough year for scientists within a community, as well as the number of forward citations these publications have garnered until today. We focus on several covariates of interest that have deeply rooted and varied theoretical foundations but very little consensus: whether breakthroughs arise at the core or periphery of the community, whether they come from scientists early or later in their careers, whether they are invented by specialists versus generalists and whether they arise from a brokerage versus a cohesive network position.

The Emergence of a Scientific Breakthrough

Scientific Breakthroughs

Breakthroughs can be depicted by various measures but its definition is ultimately linked to the notion of impact (Simonton, 1999). Breakthroughs encompass both creative novelty and success. As opposed to some inventions which become technological dead ends, breakthroughs are the foundational inventions at the basis of further incremental enhancements. In the language of the punctuated equilibrium theory, breakthroughs are the rare events that bring about discontinuities in technological trajectories, forms new technological paradigms in which their utility on the path of technological progress has been demonstrated (Dosi, 1982) and are usually associated with increased environmental turbulence (Tushman & Anderson, 1986).

The definition of *scientific* breakthrough can be constructed from the definitions of science and breakthrough. To define science, let us first draw a clear distinction between science and technology, which have been defined following two main streams. One stream, classified under the new economics of science, takes an institutional stance (Dasgupta & David, 1994;

Merton, 1957) while the other interpretation gets to the role that science and technology play respectively in generating new knowledge. Under the institutional view, science is seen as a distinctive incentive system from technology. Science although much more complex (Latour, 1987; Latour & Woolgar, 1986) than depicted herein is characterized by publication, supported by a priority-based reward system and exists predominantly, but not exclusively, in research universities. In contrast, technology is a world in which ideas are produced for economic ends and encoded in patents and other modes of protection to facilitate appropriation and commercialization (Dasgupta & David, 1994). The relationship between science and technology can also be depicted by the nature of knowledge creation. Science concentrates on demonstrating the *why* through a process of posing hypotheses that are empirically tested so as to refine theory; while technology searches for recipes for *how* by developing practical and useful techniques. In the definition of scientific breakthrough adopted herein, science is a process undertaken to understand why in a certain natural phenomenon occurs.

Following Tushman and Anderson (1986) we define scientific breakthroughs in this paper as advances that disturb the previous understandings of *why* of a particular phenomenon in a fundamental manner, give rise to novel potentially dominance upsetting technology, and have widespread technological application potential and commercialization potential.

The literature is rife with historical case studies of both scientific and technological breakthroughs. For almost every major invention, there are countless accounts written by the discoverers themselves or historians of science and technology. These accounts offer various, sometimes strongly opposing, viewpoints on the process of discovery and the interactions between winners and serious contenders. For instance, the discovery of DNA has been described at length by James Watson the winning scientist (Watson, 1963), but several other accounts are

also available notably that told from Rosalind Franklin's – another scientist in the race to determine the structure of DNA – perspective (Maddox, 2002).

These historical case studies of breakthroughs are in-depth qualitative *ex post* exposés of the history of each scientific breakthrough. Although extremely rich and incredibly insightful when characterizing the invention or discovery process, the number of stakeholders included in such historical accounts is usually limited to those in the immediate proximity of the winners, such as their mentors, collaborators and eminent fellow scientists racing for the same discovery. Consequently, these individual case studies or historical accounts lack the macro collaborative view enabled by large archival quantitative methods, and very few works have attempted to synthesize the individual findings from each. We address this gap in the literature by taking a more global view of scientists within a community using quantitative econometric analysis.

Emergence of Scientific Breakthroughs

The literature on identifying sources of breakthroughs has focused on three levels of analysis – the organizational level, the individual level and the invention level. At the organizational level, works have either concentrated on identifying organizational sources of breakthrough, in other words, revealing the types of firms more likely to invent technological breakthroughs (Fleming, 2002), or focused and contrasted the likelihood of technological breakthrough creation by entrants versus incumbents (Ahuja & Lampert, 2001) as well as their subsequent impact (Trajtenberg, 1990). These studies cite the firm's involvement in high-variance trials and merging diverse technologies as organizational sources that enhance the likelihood of breakthrough success (Fleming, 2001), and identify a curvilinear relationship

between a firm's exploration in novel, emerging and pioneering technologies and breakthrough creation (Ahuja & Lampert, 2001).

Works focusing on knowledge creation at the invention or innovation level of analysis has emphasized the process of recombinant search. Assuming bounded rationality inventors tend to search locally for an optimal recombination of existing components to create knowledge. For instance, at the heart of Henderson and Clark's (1990) architectural innovation is the notion that new innovations are created by merely rearranging the way in which the components of a product is linked leaving untouched the core design concepts of each initial component. Even though inventors tend to search locally, uncertainty still predominates the recombination process and the sources of this uncertainty is derived from the search for unfamiliar new components and new combinations (Fleming, 2001). At the individual level, studies have explored how direct relationships between inventors or scientists and their network structure impact knowledge creation. The findings point to diminishing returns to both the number of relationships as well as the frequency of relationships to knowledge creation (McFadyen & Cannella, 2004). These studies focus on either the process of knowledge creation through recombination or identify social capital characteristics that foster knowledge productivity, but they do not elaborate on where the knowledge comes from within a community of inventors or scientists – and if the sources of that future knowledge can be predicted. This gap in literature thus constitutes the first point of departure for our current paper.

The inquiry of where a scientific breakthrough occurs within a community of scientists gets to the issue of knowledge creation and novelty. Previous research in the area of knowledge creation in both science and technology at the individual researcher or inventor unit of analysis has used similar research designs in that they employed network analysis of large archival

datasets to econometrically infer relationships. In the scientific realm, the influence of collaborative networks on scientific productivity has been explored at the individual scientist level where a curvilinear relationship was established between social capital and knowledge creation (McFadyen & Cannella, 2004). Similar effects were also found using strength of relations as the explanatory variable. Cultivating relationships increases the amount of information and resources received from others, but maintaining an over abundant number of relationships soon becomes costly and outweighs its advantage due to increased startup cost and opportunity cost of time spent maintaining them (Zucker, Darby, Brewer, & Peng, 1996). With increased frequency of interactions, exchanges become more efficient but may give rise to convergence of understanding and ideas (Coleman, 1988), particularly if such interactions are with the same people. In a follow-up paper exploring the determinants of knowledge creation in science, two network measures – average tie strength and ego network density – and their interactions are studied (McFadyen, Semadeni, & Cannella, 2009). Using the same research setting as previously described, the authors find that a sparse network with low degrees of connectivity providing more opportunities for new information exchange and diverse perspectives coupled with strong ties that facilitates the transfer of tacit knowledge grants the best condition for knowledge creation.

Even though not set in the scientific realm, results from several works that studied the effect of various network measures on knowledge creation can be readily extrapolated and applied onto scientific knowledge creation. Within technology, works on the creation of knowledge focuses on determining factors influencing innovation productivity. Analysis at the firm level have studied the effect of network measures – such as collaborative networks (more specifically direct and indirect ties) and structural holes (Ahuja, 2000), as well as regional

agglomeration and network centrality (Whittington, Owen-Smith, & Powell, 2009) – on organizational innovation output. Furthermore, the effects of brokerage (Burt, 2004) and cohesion (Obstfeld, 2005; Uzzi, 1997) on knowledge creation have been studied with conflicting results as proponents of both collaborative structure camps have argued for its positive effect on knowledge productivity. Fleming, Mingo and Chen (2007) reconcile the conflict by adding another dimension – personal attributes of collaborators – which was ignored in previous works.

All above results are critical in understanding the factors influencing novel knowledge creation but they do not address the questions of where a breakthrough comes from and who makes the actual discovery; in other words, whether they come from scientists early or later in their careers, whether they are invented by specialists versus generalists and whether they arise from a brokerage versus a cohesive network position. The following sections of this paper address these questions using quantitative using network visualization tools and regression models set in a case study of the RNA interference breakthrough.

The Case of RNA Interference as a Scientific Breakthrough

RNA interference is a good candidate of such a scientific breakthrough described above because of its research and therapeutic potential. The phenomenon was initially observed by plant biologists in the early 1990s where an attempt to transgenically alter color pigmentation in petunia plants yielded unexpected outcomes. According to existing knowledge at the time, researchers were expecting darker purple flowers due to overexpressed genes from artificially introducing chalcone synthase, a flower pigmentation enzyme (Napoli, Lemieux, & Jorgensen, 1990). Instead, petunia flowers became less pigmented than their natural form, producing fully or partially white flowers which indicated to scientists that as opposed to the intended gene

overexpression, the activity of chalcone synthase had significantly decreased. This phenomenon was also found by scientists studying fungi (Romano & Macino, 1992) and was named quelling, although it was not immediately recognized as related to the phenomenon of co-suppression of gene expression found in petunia plants. Despite several early observations of the phenomenon, the underlying molecular mechanism remained unknown. One reason behind the difficulty of explaining the observed phenomenon and identifying the causal agent was the disparity in causal pathways between the RNAi mechanism and the central dogma of molecular biology. The central dogma of molecular biology is a framework that characterizes the main process of sequential genetic information flow and dictates gene expression in which information contained in double-stranded DNA is transcribed into a newly formed single stranded messenger RNA (mRNA) and subsequently translated into proteins or enzymes. Thus, in the central dogma theory, both double-stranded DNA and single stranded RNA had salient roles for long and short term information storage respectively, while no place was left for double-stranded RNA (dsRNA) (Fire, 2007). However, it turns out that double-stranded RNA is indeed the trigger agent in RNAi and was first identified by Fire and Mello in C. elegan worms (Fire et al., 1998). Fire and Mello subsequently coined the term RNA interference to characterize the phenomenon and were awarded the Nobel Prize in Physiology and Medicine in 2006 for this notable discovery.

In short, RNA interference is a naturally occurring endogenous mechanism triggered by dsRNA precursors which are processed into small interfering RNAs (siRNA) or microRNAs (miRNA) that bind to specific other RNAs and either increase or decrease their activity, for example by preventing a messenger RNA from producing a protein which ultimately induces the silencing of specific genes (Meister & Tuschl, 2004). RNA interference is valuable as a research

tool as well as in biotechnology drug development. For instance, in research, the selective and robust effect of RNAi on gene expression as synthetic dsRNA introduced into cells can induce suppression of specific genes of interest both in vitro and in vivo. It can also be applied to large-scale screenings that systematically shut down each gene in the cell, which can help identify components necessary for a particular cellular process or event. Thus exploitation of the RNA interference pathway is a promising tool in biotechnology and medicine where we can conceivably use this mechanism to treat genetic diseases by turning off Huntington's disease or certain liver cancers for example. Applying the definition of scientific breakthrough as discussed earlier, RNAi constitutes such a breakthrough because it disrupts the previous understanding of *why* of the central dogma in molecular biology through the introduction of a novel RNA interference pathway, gives rise to a new technology in silencing genes, and has widespread technological application and commercialization potential as described above. Furthermore, the naming of siRNA, a class of dsRNA involved in the RNAi pathway, as breakthrough of the year in 2002 by *Science* (Couzin, Enserink, & Service, 2002) supports our decision to research RNA interference as a scientific breakthrough case study.

Identifying a scientific community

We must first define a scientific community before answering who in that community is most likely to discover a breakthrough. How a community is defined is crucial to understanding how scientific breakthroughs arise within that community. Communities have been studied for a long time from a network analysis viewpoint by researchers spanning the social sciences (sociologists), natural sciences (physicists), and applied sciences (computer scientists and applied mathematicians). Sociologists have focused mainly on describing characteristics of social

communities – various group organizations formed in society such as families, professional and friendship circles – through social network analysis; whereas applied physicists and computer scientists have more converged towards developing computationally efficient community detection algorithms based on notions of node similarity and partitioning. Communities are, thus, defined as groups of vertices which share some common properties and/or play similar roles within the network graph (Fortunato, 2010). Furthermore the distribution of edges is locally and globally inhomogeneous, with high concentrations of edges within special groups of vertices, and low concentrations between these groups (Girvan & Newman, 2002).

The literature on community detection in physics and applied mathematics is young but plentiful. The topic of community detection gained significant traction since 2002 after Michelle Girvan and Mark Newman (2002) brought graph-partitioning problems to the attention of the broader fields of statistical physics and applied mathematics. Community detection has striking similarities with certain physics models and mathematical algorithms which explains its increasing popularity as a research interest in these fields. Numerous community detection methods have been developed in the past eight years by these physicists and mathematicians, with applications not only in sociology with the structural detection of social communities, but also in the natural sciences such as biological networks (Fortunato, 2010; Girvan & Newman, 2002; Porter, Onnela, & Mucha, 2009). Needless to say some community detection methods are more appropriate for social community detection than others; we will therefore focus on these methods and omit those, such as spectral partitioning, which require knowing the size of communities in advance that are only suitable for non-sociology applications in this present literature review. In fact, the main difficulty in community detection is that its ideal formulation is usually domain-specific.

Before getting into details on each community detection method one must first define a social community. Most social community detection methods are based on the definition that stemmed from Granovetter's (1973) empirical finding where links within communities tend to be strong while relationships between communities are more likely to be weak. Therefore, structurally, a community is a group of nodes densely connected to each other but sparsely connected to other dense groups in the network. Even though the existence of social communities is intuitively clear and has been studied by sociologists such as Coleman (1964), Freeman (2004) and Moody and White (2003), a rigorous definition of community structure is still fuzzy. Despite this lack of definitional rigor, community detection can be decomposed into complex interactions of two salient components, modules and hierarchies (Porter et al., 2009). Modules are single clusters of nodes, while repeatedly partitioning modules further into smaller more refined modules constitutes the process of hierarchical partitioning. Thus modules can be nested hierarchically, or there can simply be a collection of them.

Methods aimed at identifying communities can be divided into several main categories that include, non-exhaustively, clustering techniques such agglomerative and divisive methods, centrality-based techniques, local methods such as k-clique percolation, and modularity-based techniques. Traditional clustering techniques are intuitively appealing such as the agglomerative method that starts at a single node and attempts to connect similar clusters at each recomputation, or the divisive method which starts with a full graph and breaks it down into various communities. Centrality-based community detection is an example of such a divisive method in which edges are ranked based on betweeness centrality (Wasserman & Faust, 1994) and communities are formed by removing the edge with the largest value, i.e. the edge that lies on a large number of short paths between nodes or that has highest traffic. However, these clustering

methods do not allow for overlap in communities (nodes simultaneously belonging to several groups), a widespread characteristic among social communities. The k-clique (Kolaczyk, 2009) percolation method addresses this problem and enables overlapped community detection. A community is formed as the union of k-clique subgraphs. K-cliques represent the most-connected modules in a network representation and therefore using k-clique may cause one to overlook other dense modules that are not necessarily as well connected (Palla, Derenyi, Farkas, & Vicsek, 2005). Modularity optimization uses the modularity quality function to quantify communities. The quality function compares the number of intracommunity edges to those expected had it been generated randomly. Essentially the quality function is a measure of how well given network partitions classifies its communities. One constraint in the original formulation of modularity-based techniques is the resolution limit. Hence communities smaller than a certain threshold are undetectable as they tend to be merged with larger communities (Fortunato & Barthelemy, 2007). To address this issue researchers have opted to explicitly define a resolution parameter so as to control coarseness in the communities detected (Reichardt & Bornholdt, 2006).

Earlier techniques of detection of communities has mainly concentrated on identifying existing mature communities, other than for some notable exceptions (Hopcroft, Khan, Kulis, & Selman, 2004), and paid less attention on dynamically tracking how communities evolve from birth to growth to, finally, death. Recent papers have introduced methods with the capability to not only dynamically identify community structure across time frames but also networks that exhibit multiplicity or multiple types of links (univariate or bivariate) and have multiple scales (resolutions) (Mucha, Richardson, Macon, Porter, & Onnela, 2010). The authors combine

quality function and resolution parameter techniques as discussed above to achieve community detection across time-dependent, multiscale and multiplex networks.

A new branch of literature, that departs from this traditional structural community formed by detecting cohesive nodes described above, focus on groups of links instead (Ahn, Bagrow, & Lehmann, 2010). This novel and unorthodox method of identifying communities is able to incorporate overlapping nodes while revealing organizational hierarchy. It successfully reconciles the conflicting notions of overlap and hierarchy, but still suffers from the main limitation associated with structurally detected communities as discussed below.

Despite a fairly advanced literature on community detection, spearheaded by network physicists and applied mathematicians, as evidenced by the number of extensive reviews available in the literature (Fortunato, 2010; Porter et al., 2009), several unique and defining characteristics of our data and study still make it difficult to detect communities in a straight-forward fashion and, consequently, expose the limitations of currently available methods. Indeed, one important limitation of these structurally detected communities, which follows from the starting definition of a community consisting of cohesive group of nodes or links, is that all members of a given community must be connected to one another. "Sub" communities can be identified within a larger connected component, however, the analysis begins with – and assumes – a connected component. Thus, these methods of community detection preclude communities that share similar functional attributes yet have some unconnected members. For instance, in the case of our RNAi scientists, given our network depiction of scientific collaboration in which nodes are individual scientists and edges are co-authorship relationships, even though scientists A and B do not necessarily have a collaborative relationship reflected through co-authored publishing both work in advancing the scientific understanding of RNAi. Consequently, we

envision our community of RNAi focused scientists to be depicted as a collaboration network with several unconnected components. At this point, unfortunately, most theoretical research into community detection techniques have solely focused on structurally identifying communities, promptly ignoring nodal characteristics that define functional communities.

One potential way to incorporate both functional and structural characteristics in our definition of RNAi community is to employ a multiplex method in the jargon of network community scientist scholars. In the functional plane, because our study is centered on the period prior to breakthrough in 1998 and a set community of RNAi researchers has yet to emerge, we made use of MeSH (Medical Subject Headings) keywords to search for papers publishing in a specific area rather than a keyword search of academic publications titles and abstracts. MeSH keywords are believed to be a relatively objective classification scheme; instead of being assigned by authors themselves, MeSH is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences which can also serve as a thesaurus that facilitates searching. It is created and updated by the United States National Library of Medicine (NLM) and used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings. Consequently, we define our community of researchers at risk of discovering a breakthrough in the period prior to 1998 from their published peer-reviewed articles using the MeSH search terms[1] "RNA, Double-Stranded", "RNA, Antisense" and "Gene Expression Regulation" in PubMed. These three MeSH terms were identified following a review of articles on the history of RNA interference as well as Nobel Prize lectures. Scientists were attempting to explain gene expression regulation by experimenting with both double-stranded RNA and antisense RNA as causal agents. It is interesting to note that for that time period, "RNA,

[1] The exact search string used in PubMed query extracted on March 7, 2010: ("RNA, Double-Stranded"[Mesh] OR "RNA, Antisense"[Mesh]) AND "Gene Expression Regulation"[Mesh] AND eng[lang] AND 1978:1998[dp].

Interference" did not exist yet as a MeSH term. It was only until 2002 that it became part of the MeSH term lexicon.[2] In this way, we have effectively identified a community of scientists focused in RNAi research topically based on their research publication focus. Using this sample of patents, we then construct a collaborative network of RNAi scientists with individual scientists depicted as nodes and their co-authorship relationships representing edges linking each node. This network community, therefore, boasts a multiple-component structure instead of being all interlinked in a single large one and consists of scientists interested in similar topics.

Furthermore as the analysis focuses on the breakthrough, the baseline year is 1998 – the year that Fire and Mello published their RNAi mechanism discovery in Nature. We segment the case of RNAi for visualization to the period pre-breakthrough (1978-1998), where Figure 1 illustrates the overlayed network of scientists and inventors for the RNAi community and all their co-authors up to 1997, the year prior to breakthrough discovery. Data for the network diagrams, plotted using the iGraph package in Python 2.6, is obtained by aggregating patent (Lai, D'Amour, & Fleming, 2009) and paper (Torvik & Smalheiser, 2009) database searches, and disambiguating each author. Red nodes illustrate scientists who only publish papers, blue nodes indicate scientists that also publish patents while green nodes are those that only publish patents, and edges represent co-authorship relationships. Node size shows the productivity as measured by the number of publications, either papers or patents, each researcher has. It is interesting to note that the RNAi breakthrough by Fire and Mello did not come from the largest component, but rather arose from the periphery of the network at the 26th largest component. Due to the minimal number of patents in the pre-breakthrough period in which our present paper focuses on, we will solely employ the MedLine scientific paper database in our subsequent quantitative analyses.

[2] Please see the appendix for a detailed description of our approach.

[Insert Figure 1 about here]

Quantitative Modeling

Data and Methods

Having structurally visualized the evolution of collaborative networks after the emergence of a breakthrough in the previous section, this part of the paper explores quantitatively how the breakthrough occurred from within our defined community of scientists. The analysis is undertaken at the individual level because innovators are a critical input and at the very heart of the process of innovation in both science and technology. Given the scientists' previous personal attributes prior to 1998, we would like to observe who will have a fruitful year in 1998 thereby identifying individual characteristics that make them most likely to discover a breakthrough. These personal attributes include factors that characterize their past and present experiences as well as features of their environment, such as the age of scientists, publication history, social structure, organizational affiliation at the time of breakthrough, as well as past affiliations.

We restrict the dataset used in the empirical analysis to the PubMed Author-ity database (Torvik & Smalheiser, 2009) due to the scientific nature of the breakthrough, and organize each data point as unique author records in which information such as the number of citations, the number of papers, the prior career, and the affiliation for each author can be tabulated. Scientists defining the RNAi community pre-breakthrough is obtained using the same sample of papers from the MeSH keyword search performed in the visualization section, which yielded 789 publications and 2,546 authors. We drop from the analysis data, records without any authors, records with missing data as well as those whose career ended before 1998, which leaves us with

1,886 author records. Moreover, as our quantitative analysis focuses on observing publication performance in 1998, the explanatory variables include publication data from our sample of scientists up to and including 1997, while publication data in 1998 are used as our outcome variables.

Regression Models

Two descriptive regression model groups that mirror the nature of our dependent variables are used in the analysis. Both groups are count models with the number of publications in 1998 and forward citation counts as the outcome variable respectively. The regressions use quasi maximum likelihood Poisson models as publications and citations are non-negative counts and over-dispersed. The over-dispersion in dependent variables prevents the use of standard Poisson models in which it is assumed that the mean and variance of the variable distribution are equal.

Dependent Variables

Three outcome variables are constructed to measure publication and publication quality. The first variable is a dummy variable that simply indicates whether the scientist in our sample has published an academic paper in 1998. The remaining two dependent variables measure publication quantity and impact, and consist of the number of publications in 1998 as well as the number of forward citations these 1998 publications garner until today. Measuring creative impact requires the disentanglement of novelty and creative success. Consequently, our measure of publication impact is based on a social definition of creative success, where inventors and scientists are only thought to be creative until they receive recognition from their community or

society as a whole and use their work as a foundation to make further advancements (Simonton, 1999).

Explanatory Variables

We focus on several covariates of interest that have deeply rooted theoretical foundations, but very little consensus. Therefore, we attempt to empirically validate and shed light on these conflicting determinants of breakthrough creation from a scientific community's point of view.

Periphery vs. Core. The preliminary visualization result shows that the Fire and Mello RNAi breakthrough arises at the periphery of the community. We attempt in this section to econometrically verify our visual finding. The sociology of science literature supports a view in which successful problem-solvers may not necessarily be at the core of the problem field. The main theoretical reasoning behind this line of work is that scientists situated at the margins of their community possess "focused naïveté" – a useful ignorance of prevailing assumptions and theories (Gieryn & Hirsh, 1984). Consequently, they have access to differing knowledge and perspectives than the actors at the core. This advantage ultimately helps them in uncovering potentially novel and highly impactful breakthroughs (Gieryn & Hirsh, 1983, 1984; Handberg, 1984; Simonton, 1984). These arguments are echoed in the organizational literature which argues that breakthroughs come from "outside" an extant industry (Tushman and Anderson, 1986).

An opposing viewpoint stems from the more recent social networks literature which believes that individuals situated at the core get more influx and faster flow of information from

social ties. As knowledge creation is viewed as a recombinant search process, scientists placed at the core of their collaborative network have better access to relevant information, more resources, and are less isolated, which in turn increases their likelihood of creating breakthrough work.

An even more nuanced argument comes from sociology that argues that deviance is most likely from the core or the periphery (Damon & Zuckerman, 2001). Such a "middle status conformity" argument proposes that people who are in the core can afford to experiment, or even set the trend, in opposition to accepted convention. At the other extreme, people with little investment, or outsiders, have nothing to lose from deviating from convention. Both groups would be more likely to experiment, and assumedly invent a breakthrough. In contrast, people in the middle are most concerned about appearances and status and perceive little opportunity to experiment. Hence, they would be expected to be least likely to invent a breakthrough.

We create two measures of core mirroring both topical and structural communities that we constructed. The first measure relates to the collaborative core of the scientific network, in which the periphery/core variable is a distance measure constructed by calculating *1/number of node*s within each network component. The largest component at the core contains the most number of nodes, and hence as ones moves further away from the largest core component our distance measure of core increases since the number of nodes in each component decreases. If we used a simple measure indicating in which linked component the author is situated, with the largest component being component #1 and the smaller components having bigger component numbers, i.e. the closer to the core an author is the lower her component number will be, we run into the problem in which multiple components with the same number of nodes are randomly numbered sequentially by the network analytic software. Furthermore, using a distance measure

avoids the subjective decision associated with what defines core versus periphery if we were to use a dummy variable indicating core for the nth largest components. This structural core measure is depicted as the `collabcore` variable in the regression models.

The second measure depicts core versus periphery from a technical standpoint. Following the topical construction of the scientific community working on suppressing gene expression using MeSH keywords, technical core is calculated by tabulating the frequency of MeSH keywords "RNA, Double-Stranded", "RNA, Antisense", and "Gene Expression Regulation" and all previous variants[3] in a scientist's publication history and normalizing by the total frequency of all MeSH keywords associated with a particular scientist, i.e. $\frac{Mesh\ freq\ of\ RNA,Double-stranded+RNA,Antisense+Gene\ Expression\ Regulation}{\sum_i freq\ of\ MeSH_i}$. The more a scientist's work is focused in the key antecedent fields to RNA interference as reflected by the frequency in which their published works are classified under the above three MeSH keywords, the more they are embedded in the technical core of the community. This technical core measure is depicted as the `techcore` variable in the regression models.

Specialist vs. Generalist. Similar to the periphery versus core argument, we attempt to identify whether breakthroughs tend to come from specialists or generalists. According to advocates of marginality, a generalist is not bound to the current thinking in the focal field and can therefore offer different perspectives and heuristics that will drastically increase the probability of discovering a breakthrough (Jeppesen & Lakhani, 2010). Whereas experts deeply rooted in their respective scientific domains may suffer from a curse of knowledge that limits their exploration beyond their immediate knowledge neighborhoods.

[3] Prior MeSH keywords for "Gene Expression Regulation" include "Gene Expression", "Genes" and "Phenotype". When tabulating frequency for "Gene Expression Regulation" we also incorporated counts of its prior keywords.

On the other hand, specialists might be better positioned to solve a breakthrough because their deep knowledge in a field will enable them to optimally recombine components at their disposal. Even if their search for components is limited to local maxima they are able to make better use of these components as their expertise enables them to find the optimal breakthrough solution in their given field.

We capture the degree of expertise of each individual scientist using a publication cohesiveness measure that we implement based on the breadth of MeSH keywords in a scientist's publications. This metric is a measure of the prominence of high-frequency peaks in the unique list of MeSH keyword distribution associated with every publishing author. We first identity the top most frequent *k* number of MeSH terms for each scientist and calculate publication cohesiveness as $\frac{sum\ of\ MeSH\ freq\ in\ range\ 2\ to\ k+1}{(sum\ of\ MeSH\ freq\ in\ range\ 2\ to\ k+1)+sum\ of\ remaining\ MeSH\ freq}$ (Swanson, Smalheiser, & Torvik, 2006). This cohesiveness measure is labeled as the `pubcoh` variable in our models. According to the measure, a specialist with a narrow range of MeSH keywords with extremely high frequencies for the top ones will have a high value in the numerator, and consequently have higher cohesiveness values; whereas a generalist tends to be characterized by a more uniform MeSH keyword frequency distribution with higher variance and less defined high-frequency peaks which translates into lower numerator and cohesiveness values. The more cohesive a scientist's set of publications, the narrower their breadth of publication and the more expertise they possess in a given field.

Lifecycle. Whether a younger versus older researcher will be more at risk of breakthrough has also been debated in the literature. On the one hand, literature on the burden of knowledge (Jones, 2009; Wuchty, Jones, & Uzzi, 2007) has been developed based on the observation that

innovators are not born at the cutting edge frontier of knowledge and must undertake significant education. Furthermore, significant increase of the total stock of knowledge over the past few centuries implies that the amount of education innovators must accumulate also increases proportionally. Extending Newton's imagery of standing on the shoulders of giants, "one must first climb up their backs, and the greater the body of knowledge, the harder this climb becomes" (Jones, 2009). Innovation increases the stock of knowledge; but in order to contribute and create new knowledge, innovators must first surmount the educational burden of knowledge so as to place themselves at the frontier of science in a position with the highest probability of adding to the stock. To compensate for this ever increasing burden of knowledge, innovators and scientists have followed two paths: to learn more and/or to narrow their expertise. The implication of more learning is a delayed contribution to the stock of knowledge thus pushing back the age of first contribution (Jones, 2009).

On the other hand from a cognitive viewpoint, Simonton has studied over the past few decades the relationship between age and creativity in numerous artistic and scientific fields (Simonton, 1989). Although fields differ significantly across optimal creativity age, younger scientists were found not to be afraid to tackle hard problems, and are less weighted down or encultured with conventional wisdom. They have had less time to socialize into the norms of established institutions and can therefore freely think outside the box increasing their propensity of generating breakthroughs.

From a regression modeling standpoint, construction of scientists' `experience` is proxied by the number of years since the year of their first publication.

Brokerage vs. Cohesion. Brokerage versus cohesive collaborative structures has also been extensively studied both in terms of knowledge creation as well as diffusion. Brokerage was found to enhance innovation creativity and output since brokers occupy a nexus position in which diverse information flow to them first, thus enabling control over information and providing the best opportunity to generate new knowledge combinations (Burt, 2004). This is not to say that scientists in cohesive collaborative networks are less likely to produce innovation. In fact, proponents of cohesion argue for the benefits of trust, redundant information paths that facilitates tacit knowledge transfer, shared risk taking, and easier mobilization (Obstfeld, 2005; Uzzi, 1997). Reconciling the conflict between the brokerage and cohesion camps, Fleming, Mingo and Chen (2007) found that cohesion coupled with a researcher's personal attributes such as wider breadth of publication and various prior careers increase innovation productivity as the diversity of personal experience counteracts the staleness of cohesive networks. Furthermore, a more cohesive collaborative structure facilitates understanding of all components of the new knowledge, fosters a greater sense of mutual ownership of the creative product which increases the likeliness of the creation from being used again (Obstfeld, 2005; Uzzi, 1997).

In terms of variable construction, network cohesion, `constraint`, is calculated using Burt's constraint (Burt, 2004), and the number of prior professional affiliations, `naffil97`, is extracted from a history of affiliation data.

Control Variables

Publication History. We control for unobserved heterogeneity in the ability of scientists using publication history. The role of publications on the diffusion of innovation and knowledge has been addressed by various streams of literature in both the economics and sociology traditions.

Publications facilitate the diffusion process through disclosure of codified knowledge. A few institutional mechanisms are at work. Merton's norm of communalism (Merton, 1957) – the common ownership of scientific discoveries, according to which scientists give up intellectual property rights in exchange for recognition and esteem – combined with the winner-take-all importance accorded to scientific priority incentivizes scientists to be first at publishing their findings (Dasgupta & David, 1994). Not only do scientists gain respect from peers for publication achievements, they also obtain pecuniary recognition in the forms of renewed grants as well as promotions since publications are a critical performance evaluation criterion in universities, research institutions and science-based firms. To ascertain their primacy, scientists publish their new ideas and results to relay them to peers. And in this process, as scientists compete fiercely to be the first to discover and publish, they disseminate their ideas to the scientific community hence contributing to the overall stock of knowledge in society. This self-reinforcing mechanism between publication and scientific priority, thus, leads to the reasoning that the more publications a researcher publishes the more likely is it for their subsequent work to get reused. We control for these effects by using a count of the number of publications since first publishing until the year prior to the 1998 breakthrough, `npub97`, as well as the number of aggregated forward citations for these publications, `nforwcite97`. Since we employ the quasi maximum likelihood Poisson count model in our regressions and both variables are counts, we take their natural logarithm and denote them respectively as `lnpub97` and `lnforwcite97`.

Lone scientist vs. Team of scientists. Whether lone researchers versus teams of researchers are sources of breakthrough innovation is another frequently debated question in the innovations literature. Recent studies show a continuing and increasing trend for teams to contribute to the

production of knowledge through paper and patent publications in all natural and social science domains (Wuchty et al., 2007). Again alluding to arguments from the burden of knowledge theory, to compensate for an ever increasing body of knowledge innovators and scientists have to narrow their expertise (Jones, 2009), which translates to reduced individual capabilities which forces innovators to work more predominantly in teams. Furthermore, using patent data and moving away from the conventional analyses of the mean to incorporate tails of the citation distribution, two mechanisms in which collaboration fosters breakthrough emergence are found to be at work (Singh & Fleming, 2010). Rigorous selection processes attributed to circling ideas for critique by co-inventors decrease the likelihood of poor outcomes on the left-hand tail of the distribution, while the probability of obtaining radical outcomes is increased due to the greater opportunity in the creative search process to recombine diverse components stemming from the collectivity of collaborators.

Proponents of the lone superstar have argued that even though teams bring greater collective knowledge and effort, there are also significant costs to increased teamwork such as social network and coordination losses. Therefore a shift to teamwork may be a costly phenomenon that promotes low-impact science. However, evidence from the papers discussed above (Singh & Fleming, 2010; Wuchty et al., 2007) suggests that teams produce more highly cited work in each broad area of research. Furthermore, the citation advantage of teams has also been increasing with time when teams of fixed size are compared with solo authors.

We capture the number of co-authors each scientist collaborates with when publishing by calculating the degree network measure of each author node – number of directly linked neighboring nodes to focal node – and denote that variable using `ncoauthor97`. Following the

same reasoning as variables `lnpub97` and `lnforwcite97`, we take the natural logarithm of `ncoauthor97` when entering the variable in our regressions and denote it as `lncoauthor97`.

Affiliation. We also control for the current affiliation of researchers working in academia or corporations affect the impact of scientist's publications differently. Due to the institutional priority-based rewards system in science, higher-quality researchers may be willing to trade off more income in private firms to earn the higher expected prestige rewards in academia (Stern, 2004), especially when they are given the authority to direct their own research agendas into areas that they perceive as high-risk breakthrough areas. Thus higher-quality scientists tend to choose academia over private corporations, since researchers in academia are allowed more flexibility in pursuing their individual research agendas than in for-profit organizations. The scientist's current professional affiliation is coded from the affiliation with the most occurrences in the last ten papers she published and stored in variable `acadaffil`.

Prestige. Ranking of institutions in which scientists have been affiliated with is also an indicator of the breakthrough potential of an individual. Most university rankings are based on several criteria, one of which is the quality of research publications it produces. Thus as a researcher has been admitted to or is a faculty member of a prestigious institution, it can be assumed that they have undergone a selective admissions process which should attest to their research capabilities. Consequently, we control for the prestige of a scientist's affiliated institution by depicting the number of times the individual's past affiliations is associated with a top-tiered – top 50 – overall research university as ranked by U.S. News in 1998 and store the information in variable `prestige`.

Results

Table 1 shows the summary statistics – mean, standard deviation, minimum and maximum – for each variable used in the logit and quasi maximum likelihood, and also provides a short description of each variable. Table 2 shows the correlation matrix of covariates, which does not indicate any excessive correlations among covariates.

[Insert Table 1 about here]

[Insert Table 2 about here]

Table 3 reports the regression results for the quasi maximum likelihood Poisson models we ran using the number of 1998 publications (NPub98) as outcome variable. We first start by running a baseline model including only control variables – natural log of number of prior publications and citations garnered by these publications, type of affiliation whether academic or not, number of previous affiliations and affiliation prestige – as covariates in model 1. As expected a scientist's productivity is positively and significantly affected by academic affiliation, the number of prior publications up to 1997 she possesses, and the number of collaborators. Surprisingly the number of affiliations and the number of forward citations for pre-1998 publications negatively and significantly affect publication productivity in 1998. The slight negative relationship between the number of forward citations for pre-1998 publications and productivity in 1998 is very illustrative of the saying "quality versus quantity". We can feasibly conceive that a researcher preoccupied with publishing as many papers as possible aims for quantity at the expense of quality. Moreover, a researcher's affiliation prestige does not significantly affect 1998 publication count. This result again reinforces the quality over quantity argument, and in line with impact-based ranking methods of top-tiered institutions, prestige does not affect publication quantity but rather publication value.

In the subsequent models, we incorporate our explanatory variables of interest. We first look at their separate effects by running the baseline model plus each individual covariate in models 2 to 5. Model 2 adds Burt's constraint measure as the sole covariate and yields no significant effect on publication count. Model 3 sheds light on the specialist versus generalist debate by including the measure of publication cohesiveness which yields a negative and significant effect on the dependent variables. Thus the least concentrated the expertise of a scientist the more publications she produces in 1998. Model 4 depicts the core versus periphery arguments by incorporating both technical and collaborative measures of core. Technical core does not result in a significant relationship; however, the results pertaining to collaborative core show that the further away a scientist's collaborative network component is from the largest core component the less productive they are. Finally, model 5 looks at the lifecycle theme and finds that younger scientists are more productive.

The full model that incorporates all covariates is presented in model 6 and most of the variables significant in the previous models have maintained their significance. For instance, the effect size for a one standard deviation increase in the collaborative core measure, i.e. moving one standard deviation away from the core, is associated with a 6.7%[4] decrease in the number of publications using the coefficient from the full model. Similarly, the effect size of publication cohesion is such that a one standard deviation increase in publication cohesion decreases the number of publications in 1998 by 6.3%[5] in the full model. While a one standard deviation

[4] $\frac{e^{-1.258(\mu_{pubcoh}+\sigma_{pubcoh})}}{e^{-1.258(\mu_{pubcoh})}} = 0.933$

[5] $\frac{e^{-0.477(\mu_{collabcore}+\sigma_{collabcore})}}{e^{-0.477(\mu_{collabcore})}} = 0.937$

increase in a researcher's experience decreases publication production by 35.1%[6] in the full model.

[Insert Table 3 about here]

Table 4 depicts similar quasi maximum likelihood Poisson regression results as in Table 3, but this time using the number of forward citations to the 1998 publications (NForwCite98) as the outcome variable. The analysis format is also comparable as we start with the baseline model with control variables – natural log of number of prior publications and citations garnered by these publications, type of affiliation whether academic or not, number of previous affiliations and affiliation prestige – as covariates. Because this new set of regressions employs impact of 1998 publications as the dependent variable, we would expect a scientist's prior impact to be positively and significantly correlated, and indeed we observe this result. Similar to the baseline results in the previous group of regressions, we find forward citation counts to be positively and significantly affected by academically affiliated researchers and the number of collaborators. The number of affiliations up to 1997 and the number of pre-1998 publications negatively and significantly affect 1998 publication impact. Prestige, on the other hand, is positively and significantly correlated with 1998 publication impact. These results – positive coefficient on prestige and negative coefficient on prior publication productivity – reflect the differentiation between quality and quantity as the more prestigious top-tiered research institutions are ranked based on their research impact.

Similar to the analysis format for the count regressions with 1998 publication productivity as outcome variable, we incorporate each explanatory variables of interest independently in the models that follow. We first look at their separate effects by running the

[6] $\frac{e^{-0.0411(\mu_{experience}+\sigma_{experience})}}{e^{-0.0411(\mu_{exoerience})}} = 0.649$

baseline model plus each individual covariate in models 2 to 5. Model 2 adds Burt's constraint measure as the sole covariate and yields no significant effect on publication impact. Model 3 includes publication cohesiveness, but is inconclusive as to whether specialists or generalists are more at risk of breakthrough innovation. Again, model 4 depicts the core versus periphery arguments by incorporating both the technical and collaborative measures of core. Unlike in the productivity cases, technical core does result in a negative significant relationship with publication impact and shows that scientists at the technical periphery tend to produce more impactful papers; whereas, collaborative core has no significant effect on publication impact. Finally, model 5 looks at the lifecycle theme and, consistent with productivity results, finds that younger scientists are more productive.

The full model that incorporates all covariates is presented in model 6. Most variables that were significant in the previous models have maintained their significance. For instance, the effect size for a one standard deviation increase in the technical core measure is associated with a 14.3%[7] decrease in publication impact using the coefficient from the full model. Thus scientists at the technical periphery are more likely to publish breakthrough papers. Furthermore, the effect size of experience is such that a one standard deviation increase in experience decreases the number of publications in 1998 by 42.4%[8] in the full model.

[Insert Table 4 about here]

Discussion and Conclusion

[7] $\frac{e^{-4.939(\mu_{techcore}+\sigma_{techcore})}}{e^{-4.939(\mu_{techcore})}} = 0.857$

[8] $\frac{e^{-0.0507(\mu_{experience}+\sigma_{experience})}}{e^{-0.507(\mu_{exoerience})}} = 0.576$

The above results not only shed light on competing theories of breakthrough emergence, but also illustrate the fundamental distinction between publication productivity and impact. Referring back to the core vs. periphery, brokerage vs. cohesion, generalist vs. specialist and young versus experienced debates, our results show that they contribute differently to a researcher's productivity and subsequent impact. For instance, the generalist versus specialist debate was only conclusive for productivity but not impact, where we observed that generalists are more productive than specialists. The only consistent result we obtained was with respect to the younger versus older debate. Younger researchers are not only more productive they are also more prone for breakthrough, lending empirical evidence to Simonton's argument that younger scientists have had less time to be weighed down and encultured by the conventional wisdom of their fields (Simonton, 1989). Surprisingly, our results do not offer any clarification in the heated brokerage versus cohesion debate. Both in terms of productivity and impact, our regression results did not yield any significance.

The most interesting result stems from the core versus periphery debate. Technical periphery is a significant predictor of breakthrough innovation, whereas collaborative core is a significant predictor of knowledge production. On the one hand, scientists at the technical periphery are familiar with and have access to a broader knowledge base and more fields of study, which they can freely recombine components originating from seemingly disconnected fields. Together with the reasoning of focused naïveté, scientists at the technical periphery of their community have increased potential of uncovering highly impactful breakthroughs. On the other hand, collaborative core is associated with increased productivity and is in line with the viewpoint that individuals situated at the core enjoy more influx and faster flow of information from social ties. These results may be an empirical reflection of the middle status conformity

theory, where people at the two extremes – core and periphery – can afford to experiment in order to set a new trend in the case of the former, or have nothing to lose from deviating from the convention in the latter. Our results show that those at the collaborative core do experiment more and are more productive, but the increased experimentation does not necessarily result in breakthroughs. Conversely, those at the technical periphery are more at risk of breakthrough. Without doubt, these intriguing results warrant further detailed analysis, and the question of breakthroughs emerging from the core or periphery of a technical and/or collaborative community can be the focal research question of a subsequent paper.

Despite shedding light on several debates within the literature with regard to factors that increases the likelihood of breakthrough emergence, this paper suffers from two main limitations. First, even though we have empirically verified several theoretical debates, we still lack an in-depth understanding of the mechanism of breakthrough emergence. Thus, this work also serves a prequel to our inductive theory building paper in which we plan to interview not only the researchers who discovered the breakthrough but also those at risk, complementing the quantitative analysis so as to tell a more comprehensive story. It enables us to develop two sampling methods – residual analysis and matching methods – that determine the list of scientists to interview in our qualitative field work. Residual analysis identifies scientists who significantly underperformed or over-performed in terms of productivity and impact with respect to predictions from our regression models, while the matching method of interviewee sampling allows us to find individuals most similar to the winning scientists and understand why these matched scientists were not successful in discovering the breakthrough.

Second, the reader might wonder about the generalizability of our results given that we have explored the emergence of breakthroughs in the context of a single case study. Our focus on

the RNA interference breakthrough is partly due to the constraint of the qualitative paper in which it is difficult to interview scientists with highly impactful publications in all disciplines. As part of our future we intend to move up levels of analysis from individuals to communities and study the emergence of breakthroughs amongst many communities.

In conclusion, this paper summarizes a study that aims at understanding where a breakthrough comes from within a scientific community. It employs a mixture of quantitative regression analysis and network visualization methods to empirically answer our research question through a case study of the RNAi breakthrough.

This project is the first piece of a larger research agenda studying the co-evolution process of science and technology. The present study identifies where a breakthrough arises and its subsequent impact at the individual level of analysis. Future studies will build on the results of this project. An immediate follow-on study could explore the same research questions but focusing on teams as the level of analysis. Studies have shown that lone scientists and inventors producing significant breakthroughs are becoming more and more a myth (Singh & Fleming, 2010; Wuchty et al., 2007). Thus we can ask questions such as what mixture of characteristics in team members yield the optimal group with the highest probability of breakthrough. These characteristics can include the mixture of team members' age or tenure, disciplines, expertise depth within a field, frequency of prior collaboration, degree of knowledge overlap, etc. Another future study based in the epoch prior to breakthrough can generalize and test the theoretical findings obtained in this present article deductively at the community unit of analysis using a novel database that combines MedLine academic paper and patent databases. A generalized theory for emergence of scientific breakthroughs will inform policy decisions and help grant funding agencies decide who to sponsor.

Once a breakthrough has occurred, research can be undertaken to understand how knowledge generated in science flows into technology and conversely how technology may influence science. The existing literature on the co-evolution of science and technology is fairly thin, whereas for the most part, studies have focused on either science (McFadyen et al., 2009) or technology (Allen, 1977; Fleming et al., 2007). Those that looked at the interplay of science and technology have not explored the collaborative networks of scientists and inventors (Azoulay, Ding, & Stuart, 2009; Cockburn & Henderson, 1998). The very few that began to consider how the collaborative networks of scientists and inventors co-evolve and mutually influence each other have, however, generally been at the level of the paper-patent pair (Murray, 2002). No work has studied co-evolution at the level of analysis of the entire network (overlayed science and technology networks) for a given scientific breakthrough. None have taken a global but detailed individual researcher perspective, and compared different examples of the transfer and influence of knowledge between science and technology. Furthermore, none have temporally traced the development of a scientific breakthrough through various stages of its growth from the pre-breakthrough period, to the post-breakthrough pre-commercialization phase, and finally to the post-commercialization era.

Figure 1. The community of science researchers prior to the Mello Fire RNA interference breakthrough (1993-1998). Red nodes illustrate scientists who only publish papers, blue nodes indicate scientists that also publish patents, green that only patents, and links represent co-authorship.

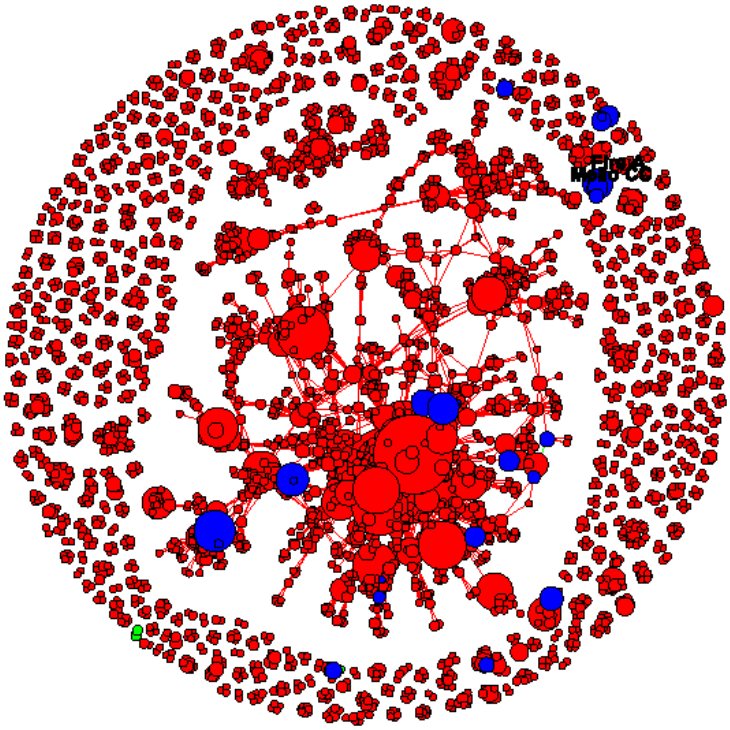

Table 1. Descriptive Statistics of Variables Used in Regression Models

Variable	Mean	Std. Dev.	Min	Max	Description
Pub98	0.83	0.38	0.00	1.00	Dummy variable indicating whether scientist published or not in 1998
npub98	3.47	5.09	0.00	87.00	Count variable for number of publications in 1998
nforwcite98	31.28	92.68	0.00	2321.00	Count variable for number of forward citations to focal 1998 publications
npub97	47.01	72.14	1.00	1181.00	Count variable for number of publications from start of career to 1997
lnpub97	3.11	1.27	0.69	7.07	Natural logarithm of count variable for number of publications from start of career to 1997
nforwcite97	474.93	1145.94	0.00	15921.00	Count variable for number of forward citations to all publication up to 1997
lnforwcite97	4.76	1.85	0.00	9.68	Natural logarithm of count variable for number of forward citations to all publication up to 1997
ncoauthor97	5.15	3.80	1.00	42.00	Count variable for number of co-authors up to 1997
lncoauthor97	1.67	0.54	0.69	3.76	Natural logarithm of count variable for number of co-authors up to 1997
acadaffil	0.99	0.11	0.00	1.00	Dummy variable indicating whether scientist published or not in 1998
naffil97	1.92	0.98	1.00	7.00	Count variable for number of affiliations associated to scientist up to 1997
prestige	0.67	0.96	0.00	5.00	Measure of affiliation institution's reputation
constraint	0.70	0.26	0.13	1.39	Network measure of Burt's constraint
pubcoh	0.47	0.06	0.30	0.80	Measure of publication breadth in a scientist's publications up to 1997
collabcore	0.16	0.14	0.00	0.50	Measure of collaborative core
techcore	0.02	0.03	0.00	0.22	Measure of technical core
experience	14.61	10.88	0.00	51.00	Number of years from first publishing to 1997

Table 2. Correlation Matrix of Covariates

		1	2	3	4	5	6	7	8	9	10	11
1	lnpub97	1										
2	lnforwcite97	0.7726	1									
3	lncoauthor97	-0.0017	-0.0687	1								
4	acadaffil	0.0316	0.035	0.0091	1							
5	naffil97	0.5307	0.4216	-0.0546	0.0519	1						
6	prestige	0.1595	0.2575	-0.084	-0.0069	0.1069	1					
7	constraint	-0.003	0.0508	-0.8969	-0.0179	0.056	0.0613	1				
8	pubcoh	0.3035	0.2404	-0.0747	0.0308	0.0902	0.0626	0.071	1			
9	collabcore	0.0364	0.0719	-0.6883	-0.0016	0.0841	0.0602	0.754	0.0394	1		
10	techcore	-0.2133	-0.1259	0.0652	-0.0252	-0.1054	-0.0335	-0.0416	-0.0741	-0.0448	1	
11	experience	0.8288	0.6284	-0.0592	0.0315	0.6504	0.0931	0.0569	0.2078	0.0912	-0.1701	1

Table 3. Quasi Maximum Likelihood Poisson Model with NPub98 as Dependent Variable (N= 1,886)

	(1) npub98	(2) npub98	(3) npub98	(4) npub98	(5) npub98	(6) npub98
npub98						
lnpub97	0.743***	0.748***	0.742***	0.750***	0.737***	0.980***
	(0.0384)	(0.0382)	(0.0382)	(0.0376)	(0.0394)	(0.0400)
lnforwcite97	-0.0322+	-0.0381+	-0.0327+	-0.0319+	-0.0296	-0.0352+
	(0.0194)	(0.0198)	(0.0194)	(0.0192)	(0.0194)	(0.0184)
lncoauthor97	0.146***	0.149***	0.110	0.134***	0.0689	0.0611
	(0.0403)	(0.0403)	(0.0808)	(0.0399)	(0.0560)	(0.0769)
acadaffil	1.396***	1.406***	1.394***	1.408***	1.393***	1.402***
	(0.277)	(0.277)	(0.277)	(0.284)	(0.279)	(0.292)
naffil97	-0.264***	-0.265***	-0.263***	-0.266***	-0.260***	-0.118***
	(0.0336)	(0.0339)	(0.0336)	(0.0329)	(0.0334)	(0.0316)
prestige		0.0296				0.00981
		(0.0226)				(0.0225)
constraint			-0.0848			0.111
			(0.179)			(0.184)
pubcoh				-1.098*		-1.258**
				(0.456)		(0.426)
collabcore					-0.493*	-0.477*
					(0.212)	(0.223)
techcore					-0.574	-0.490
					(0.865)	(0.826)
experience						-0.0411***
						(0.00409)
_cons	-2.360***	-2.383***	-2.237***	-1.852***	-2.142***	-2.053***
	(0.296)	(0.296)	(0.390)	(0.353)	(0.314)	(0.431)
N	1886	1886	1886	1886	1886	1886
ll	-4346.4	-4343.7	-4346.1	-4337.4	-4338.0	-4144.2

Standard errors in parentheses
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 4. Quasi Maximum Likelihood Poisson Model with NForwCite98 as Dependent Variable (N= 1,886)

	(1) nforwcite98	(2) nforwcite98	(3) nforwcite98	(4) nforwcite98	(5) nforwcite98	(6) nforwcite98
nforwcite98						
lnpub97	-0.286**	-0.270**	-0.286**	-0.297**	-0.304***	0.00929
	(0.0879)	(0.0844)	(0.0877)	(0.0948)	(0.0899)	(0.106)
lnforwcite97	0.816***	0.790***	0.816***	0.816***	0.825***	0.793***
	(0.0488)	(0.0438)	(0.0494)	(0.0491)	(0.0492)	(0.0458)
lncoauthor97	0.361***	0.384***	0.397*	0.378***	0.385**	0.453*
	(0.0936)	(0.0965)	(0.179)	(0.0949)	(0.140)	(0.192)
acadaffil	1.650**	1.675**	1.653**	1.628**	1.615**	1.558**
	(0.542)	(0.543)	(0.543)	(0.536)	(0.538)	(0.539)
naffil97	-0.109+	-0.114+	-0.110+	-0.105	-0.113+	0.0382
	(0.0637)	(0.0647)	(0.0637)	(0.0678)	(0.0635)	(0.0969)
prestige		0.107*				0.0995*
		(0.0462)				(0.0474)
constraint			0.0942			0.206
			(0.352)			(0.354)
pubcoh				1.413		1.331
				(2.210)		(2.079)
collabcore					-0.0313	-0.0395
					(0.462)	(0.486)
techcore					-4.867*	-4.939*
					(1.997)	(1.977)
experience						-0.0507***
						(0.0149)
_cons	-2.225***	-2.281***	-2.360**	-2.880*	-2.127***	-3.442*
	(0.585)	(0.587)	(0.793)	(1.191)	(0.631)	(1.374)
N	1886	1886	1886	1886	1886	1886
ll	-48632.0	-48239.5	-48628.0	-48507.0	-48139.3	-45256.8

Standard errors in parentheses
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

References

Adams, J. D. 1990. Fundamental Stocks of Knowledge and Productivity Growth. ***The Journal of Political Economy***, 98(4): 673-702.

Ahn, Y.-Y., Bagrow, J. P., & Lehmann, S. 2010. Link communities reveal multiscale complexity in networks. ***Nature***, 465.

Ahuja, G. 2000. Collaboration Networks, Structural Holes, and Innovation: A Longitudinal Study. ***Administrative Science Quarterly***, 45(3): 425-455.

Ahuja, G., & Lampert, C. M. 2001. Entrepreneurship in the Large Corporation: A Longitudinal Study of How Established Firms Create Breakthrough Inventions. ***Strategic Management Journal***, 22(6/7): 521-543.

Allen, T. J. 1977. ***Managing the flow of technology : technology transfer and the dissemination of technological information within the R&D organization***. Cambridge, Mass.: MIT Press].

Azoulay, P., Ding, W., & Stuart, T. 2009. The Impact of Academic Patenting on the Rate, Quality, and Direction of (Public) Research Output. ***The Journal of Industrial Economics***, 57(4): 637-676.

Burt, R. S. 2004. Structural Holes and Good Ideas. ***The American Journal of Sociology***, 110(2): 349-399.

Cockburn, I. M., & Henderson, R. M. 1998. Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery. ***The Journal of Industrial Economics***, 46(2): 157-182.

Coleman, J. S. 1964. ***An Introduction to Mathematical Sociology***. London: Collier-Macmillan.

Coleman, J. S. 1988. Social Capital in the Creation of Human Capital. ***The American Journal of Sociology***, 94: S95-S120.

Couzin, J., Enserink, M., & Service, R. F. 2002. Breakthrough of the Year: Small RNAs Make Big Splash. ***Science***, 298(5602): 2296-2303.

Damon, J. P., & Zuckerman, E. W. 2001. Middle-Status Conformity: Theoretical Restatement and Empirical Demonstration in Two Markets. ***The American Journal of Sociology***, 107(2): 379-429.

Dasgupta, P., & David, P. A. 1994. Toward a new economics of science. ***Research Policy***, 23(5): 487-521.

Dosi, G. 1982. Technological paradigms and technological trajectories : A suggested interpretation of the determinants and directions of technical change. ***Research Policy***, 11(3): 147-162.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. ***Nature***, 391(6669): 806.

Fire, A. Z. 2007. Gene silencing by double-stranded RNA. ***Cell Death & Differentiation***, 14(12): 1998-2012.

Fleming, L. 2001. Recombinant Uncertainty in Technological Search. ***Management Science***, 47(1): 117-132.

Fleming, L. 2002. Finding the organizational sources of technological breakthroughs: the story of Hewlett-Packard's thermal ink-jet. ***Industrial & Corporate Change***, 11(5): 1059-1084.

Fleming, L., & Sorenson, O. 2004. Science as a Map in Technological Search. ***Strategic Management Journal***, 25(8/9): 909-928.

Fortunato, S. 2010. Community detection in graphs. ***Physics Reports***, 486(3-5): 75-174.
Fortunato, S., & Barthelemy, M. 2007. Resolution limit in community detection. ***Proceedings of the National Academy of Sciences of the United States of America***, 104: 36-41.
Freeman, L. C. 2004. ***The Development of Social Network Analysis: A Study in the Sociology of Science***. Vancouver: Empirical Press.
Gieryn, T. F., & Hirsh, R. F. 1983. Marginality and Innovation in Science. ***Social Studies of Science***, 13(1): 87-106.
Gieryn, T. F., & Hirsh, R. F. 1984. Marginalia: Reply to Simonton and Handberg. ***Social Studies of Science***, 14(4): 624.
Girvan, M., & Newman, M. E. J. 2002. Community Structure in Social and Biological Networks. ***Proceedings of the National Academy of Sciences of the United States of America***, 99(12): 7821-7826.
Granovetter, M. S. 1973. The Strength of Weak Ties. ***The American Journal of Sociology***, 78(6): 1360-1380.
Handberg, R. 1984. Response to Gieryn and Hirsh. ***Social Studies of Science***, 14(4): 622-624.
Henderson, R. M., & Clark, K. B. 1990. Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. ***Administrative Science Quarterly***, 35(1): 9-30.
Herrera, M., Roberts, D. C., & Gulbahce, N. 2010. Mapping the evolution of scientific fields. ***PLoS ONE***, 5(5): e10355.
Hopcroft, J., Khan, O., Kulis, B., & Selman, B. 2004. Tracking evolving communities in large linked networks. ***Proceedings of the National Academy of Sciences of the United States of America***, 101(Suppl 1): 5249-5253.
Jaffe, A. B., & Trajtenberg, M. 1996. Flows of Knowledge from Universities and Federal Laboratories: Modeling the Flow of Patent Citations over Time and across Institutional and Geographic Boundaries. ***Proceedings of the National Academy of Sciences of the United States of America***, 93(23): 12671-12677.
Jeppesen, L. B., & Lakhani, K. R. 2010. Marginality and Problem Solving Effectiveness in Broadcast Research. ***Organization Science***, Forthcoming.
Jones, B. F. 2009. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? ***Review of Economic Studies***, 76(1): 283-317.
Kolaczyk, E. D. 2009. ***Statistical Analysis of Network Data: Methods and Models***. New York: Springer Science.
Lai, R., D'Amour, A., & Fleming, L. 2009. The careers and co-authorship networks of U.S. patent-holders, since 1975. ***Harvard Business School Working Paper***.
Latour, B. 1987. ***Science in action : how to follow scientists and engineers through society***. Milton Keynes ;Philadelphia: Open University Press.
Latour, B., & Woolgar, S. 1986. ***Laboratory life : the construction of scientific facts***. Princeton, N.J.: Princeton University Press.
Maddox, B. 2002. ***Rosalind Franklin : the dark lady of DNA*** (1st ed.). New York: HarperCollins.
McFadyen, M. A., & Cannella, A. A. J. 2004. Social Capital and Knowledge Creation: Diminishing Returns of the Number and Strength of Exchange Relationships ***Academy of Management Journal***, 47(5): 735-746.

McFadyen, M. A., Semadeni, M., & Cannella, J. A. A. 2009. Value of Strong Ties to Disconnected Others: Examining Knowledge Creation in Biomedicine. ***Organization Science***, 20(3): 552-564.

Meister, G., & Tuschl, T. 2004. Mechanisms of gene silencing by double-stranded RNA. ***Nature***, 431(7006): 343-349.

Merton, R. K. 1957. Priorities in Scientific Discovery: A Chapter in the Sociology of Science. ***American Sociological Review***, 22(6): 635-659.

Moody, J., & White, D. R. 2003. Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. ***American Sociological Review***, 68: 103-127.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J.-P. 2010. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. ***Science***, 328: 876-878.

Murray, F. 2002. Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. ***Research Policy***, 31(8-9): 1389-1403.

Napoli, C., Lemieux, C., & Jorgensen, R. 1990. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. ***The Plant Cell***, 2(4): 279-289.

Obstfeld, D. 2005. Social Networks, the Tertius Iungens Orientation, and Involvement in Innovation. ***Administrative Science Quarterly***, 50(1): 100-130.

Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. ***Nature***, 435: 814-818.

Porter, M. A., Onnela, J.-P., & Mucha, P. J. 2009. Communities in Networks. ***Notices of the AMS***, 56(9): 1082-1166.

Reichardt, J., & Bornholdt, S. 2006. Statistical mechanics of community detection. ***Physical Review E***, 74: 016110.

Romano, N., & Macino, G. 1992. Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. ***Molecular Microbiology***, 6(22): 3343-3353.

Scherer, F. M., & Harhoff, D. 2000. Technology policy for a world of skew-distributed outcomes. ***Research Policy***, 29(4-5): 559-566.

Schumpeter, J. A. 1942. ***Capitalism, socialism, and democracy*** (1st Harper Perennial Modern Thought ed.). New York: HarperPerennial.

Simonton, D. K. 1984. Is the Marginality Effect All That Marginal? ***Social Studies of Science***, 14(4): 621-622.

Simonton, D. K. 1989. Age and creative productivity: Nonlinear estimation of an information-processing model. ***International Journal of Aging and Human Development***, 29: 23-37.

Simonton, D. K. 1999. ***Origins of genius : Darwinian perspectives on creativity***. New York: Oxford University Press.

Singh, J., & Fleming, L. 2010. Lone Inventors as Sources of Breakthroughs: Myth or Reality? ***Management Science***, 56(1): 41-56.

Stern, S. 2004. Do Scientists Pay to Be Scientists? ***Management Science***, 50(6): 835-853.

Swanson, D. R., Smalheiser, N. R., & Torvik, V. I. 2006. Ranking indirect connections in literature-based discovery: The role of medical subject headings. ***Journal of the American Society for Information Science & Technology***, 57(11): 1427-1439.

Torvik, V. I., & Smalheiser, N. R. 2009. Author Name Disambiguation in MEDLINE. ***ACM transactions on knowledge discovery from data***, 3(3): 1-29.

Trajtenberg, M. 1990. A Penny for Your Quotes: Patent Citations and the Value of Innovations. ***The RAND Journal of Economics***, 21(1): 172-187.

Tushman, M. L., & Anderson, P. 1986. Technological Discontinuities and Organizational Environments. ***Administrative Science Quarterly***, 31(3): 439-465.

Uzzi, B. 1997. Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness. ***Administrative Science Quarterly***, 42(1): 35-67.

Wasserman, S., & Faust, K. 1994. Social Network Analysis: Methods and Application, ***Structural Analysis in the Social Sciences***. Cambridge: Cambridge University Press.

Watson, J. D. 1963. ***DNA : the secret of life*** (1st ed.). New York: Alfred A. Knopf.

Whittington, K. B., Owen-Smith, J., & Powell, W. W. 2009. Networks, Propinquity, and Innovation in Knowledge-intensive Industries. ***Administrative Science Quarterly***, 54(1): 90-122.

Wuchty, S., Jones, B. F., & Uzzi, B. 2007. The Increasing Dominance of Teams in Production of Knowledge. ***Science***, 316(5827): 1036-1039.

Zucker, L. G., Darby, M. R., Brewer, D. D., & Peng, Y. 1996. ***Collaboration Structure and Information Dilemmas in Biotechnology***. Thousand Oaks, Calif.: Sage.