A CASE STUDY ON

"Global Innovation Network and Analysis (GINA)"

SUBMITTED TOWARDS THE
PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

BACHELOR OF ENGINEERING (TE Computer Engineering)

BY

Anushka Dhamdhere	TEB03
Aditya More	TEB04
Harshada Kulkarni	TEB08
Ritesh Bhokare	TEB10

Under The Guidance of
Prof. Devyani J. Bonde

"Towards Ubiquitous Computing Technology"
DEPARTMENT OF COMPUTER ENGINEERING
Marathwada Mitra Mandal's
Institute of Technology (MMIT)
Lohgaon, Pune- 411 047

(2023-24)

“येथे बहुतांचे हित”
MMIT

“Techno-Social Excellence”
Marathwada Mitra Mandal’s
Institute of Technology (MMIT)
Lohgaon, Pune- 411 047
Accredited ‘A’ Grade by NAAC

“Towards Ubiquitous Computing Technology”
DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project Entitled
“Global Innovation Network and Analysis (GINA)”

Submitted by

Anushka Dhamdhere	TEB03
Aditya More	TEB04
Harshada Kulkarni	TEB08
Ritesh Bhokare	TEB10

is a bonafide work carried out by students under the supervision of Prof. Devyani J. Bonde and it is submitted towards the partial fulfilment of the requirement of Bachelor of Engineering (TE Computer Engineering) DSBDA Laboratory.

Prof. Devyani J. Bonde Internal Guide Dept. of Computer Engg.	Dr. Subhash G. Rathod H.O.D Dept. of Computer Engg.

Abstract

Recent literature developments in the realm of global firm networks underscore the significance of not just global production networks (GPNs) but also global innovation networks (GINs). Nevertheless, there remains a lack of theoretical and empirical clarity regarding the distinctions, similarities, and interactions between firms' GINs and GPNs. This paper endeavors to address this gap by employing case studies and social network analysis to delineate the network characteristics of two case study firms' GINs and GPNs.

The swift transition of China and India from low-cost producers to innovators has spurred heightened interest in the globalization of innovation activities, particularly the proliferation of GINs. However, existing literature primarily consists of either theoretical discussions or limited case studies. Consequently, there is a lack of understanding regarding the diverse forms of GINs in which firms engage, encompassing various levels of global reach, innovativeness, network integration, and other key attributes.

To address this gap, this paper proposes a firm-centric taxonomy of global innovation networks that accounts for these diverse dimensions. Drawing on survey-based firm-level data from five European countries, as well as Brazil, China, India, and South Africa, it furnishes empirical insights into the characteristics of different variants of global innovation networks. This endeavor represents the first comprehensive theoretical and empirical examination of the manifold forms of global innovation networks, contributing significantly to the literature.

Global Innovation Network and Analysis (GINA)

Introduction

EMC's Global Innovation Network and Analytics (GINA) team comprises seasoned technologists situated in centers of excellence (COEs) across the globe. Tasked with fostering innovation, conducting research, and nurturing university partnerships, the team underwent a transformation in 2012 under a newly appointed director. The aim was to enhance these endeavors and establish a framework for tracking and analyzing relevant information. Additionally, the team sought to devise more robust mechanisms for capturing the outcomes of informal dialogues with EMC's thought leaders, academic circles, and other organizations, envisioning these insights as valuable resources for future exploration.

With a belief that their approach could facilitate global idea exchange and enhance knowledge sharing among geographically dispersed GINA members, the team devised a plan centered around three main objectives:

1. Establishing a data repository to house both structured and unstructured data.
2. Monitoring research activities conducted by technologists worldwide.
3. Extracting patterns and insights from the accumulated data to enhance the team's operational efficiency and strategic direction.

The GINA case study serves as a testament to how the team leveraged the Data Analytics Lifecycle to scrutinize innovation data at EMC. Given the inherent complexity of measuring innovation, the team aimed to employ advanced analytical techniques to pinpoint key innovators within the organization.

Phase 1: Discovery

During the discovery phase of the GINA project, the team initiated the process by identifying potential data sources. While GINA boasted a team of technologists skilled in various engineering domains, it lacked a formal analytics team to execute the envisioned analyses. Seeking guidance, the team consulted experts such as Tom Davenport from Babson College, renowned for his expertise in analytics, and Peter Gloor from MIT, a specialist in collective intelligence and creator of CoIN (Collaborative Innovation Networks). Subsequently, the team opted to crowdsource the analytics work by enlisting volunteers from within EMC.

Roles within the working team were allocated as follows:

- Business User, Project Sponsor, Project Manager: Vice President from the Office of the CTO
- Business Intelligence Analyst: Representatives from IT
- Data Engineer and Database Administrator (DBA): Representatives from IT
- Data Scientist: Distinguished Engineer, also responsible for developing the social graphs depicted in the GINA case study

The project sponsor adopted a strategy centered around leveraging social media and blogging to expedite the collection of innovation and research data globally and to motivate teams of "volunteer" data scientists across various locations within EMC. Despite lacking a formal team, the sponsor demonstrated resourcefulness in identifying individuals capable and willing to contribute their time to tackle intriguing problems. Harnessing the passion of highly skilled individuals for data, the sponsor managed to engage them in creatively addressing challenging tasks.

Data for the project were categorized into two main types:

A. Five years' worth of idea submissions from EMC's internal innovation contests, known as the Innovation Roadmap (formerly the Innovation Showcase). This data comprised both structured elements (such as idea counts, submission dates, and inventor names) and unstructured content (the textual descriptions of ideas).

B. Minutes and notes documenting innovation and research activities worldwide. This data also included a mix of structured attributes (dates, names, geographic locations) and unstructured information providing rich insights into knowledge growth and transfer within the company. Often, such information resides in business silos with limited visibility across disparate research teams.

The GINA team formulated ten main Innovation Hypotheses (IHs), encompassing:

1. Mapping innovation activity to corporate strategic directions across different regions.
2. Examining the impact of global knowledge transfer on the time required to deliver ideas.

3. Assessing the relationship between participation in global knowledge transfer and idea delivery speed.
4. Analyzing idea submissions to predict funding likelihood.
5. Measuring and comparing knowledge discovery and growth across geographic regions.
6. Identifying research-specific boundary spanners through knowledge transfer activities.
7. Mapping strategic corporate themes to geographic regions.
8. Evaluating the effect of frequent knowledge expansion and transfer events on asset generation speed.
9. Using lineage maps to identify instances where knowledge expansion and transfer did not result in a corporate asset.
10. Classifying emerging research topics and mapping them to specific individuals, boundary spanners, and assets.

These IHs were grouped into two categories:

- Descriptive analytics: Understanding current trends to foster creativity, collaboration, and asset generation.
- Predictive analytics: Providing guidance to executive management on future investment directions.

Phase 2: Data Preparation

The team partnered with its IT department to set up a new analytics sandbox to store and experiment on the data. During the data exploration exercise, the data scientists and data engineers began to notice that certain data needed conditioning and normalization. In addition, the team realized that several missing datasets were critical to testing some of the analytic hypotheses. As the team explored the data, it quickly realized that if it did not have data of sufficient quality or could not get good quality data, it would not be able to perform the subsequent steps in the lifecycle process. As a result, it was important to determine what level of data quality and cleanliness was sufficient for the project being undertaken. In the case of the GINA, the team discovered that many of the names of the researchers and people interacting with the universities were misspelled or had leading and trailing spaces in the datastore. Seemingly small problems such as these in the data had to be addressed in this phase to enable better analysis and data aggregation in subsequent phases.

Phase 3: Model Planning

In the GINA project, leveraging social network analysis techniques to examine the networks of innovators within EMC appeared feasible for much of the dataset. However, in some instances, devising appropriate methods to test hypotheses proved challenging due to data limitations. For instance, in the case of IH9, the team opted to embark on a longitudinal study to track data points over time regarding individuals developing new intellectual property. This data collection initiative aimed to enable future testing of the following two ideas:

- IH8: Regular knowledge expansion and transfer events reduce the time required to transform an idea into a corporate asset.
- IH9: Lineage maps can unveil instances where knowledge expansion and transfer fail to result in a corporate asset.

For the proposed longitudinal study, establishing goal criteria was essential. The team needed to define the ultimate objective of a successful idea that had traversed the entire journey. Key considerations for the scope of the study included:

- Identifying pertinent milestones toward achieving the goal.
- Tracing how individuals progress ideas from each milestone toward the ultimate objective.
- Comparing the trajectories of ideas that succeed with those that fail to reach the goal.
- Assessing timeframes and outcomes using various methodologies, including t-tests or classification algorithms, depending on the data collection and assembly methods employed.

Phase 4: Model Building

The GINA team utilized various analytical methods, including Natural Language Processing (NLP) techniques applied by the data scientist to analyze textual descriptions from the Innovation Roadmap ideas. Additionally, social network analysis was conducted using R and RStudio, followed by the development of social graphs and visualizations using R's ggplot2 package.

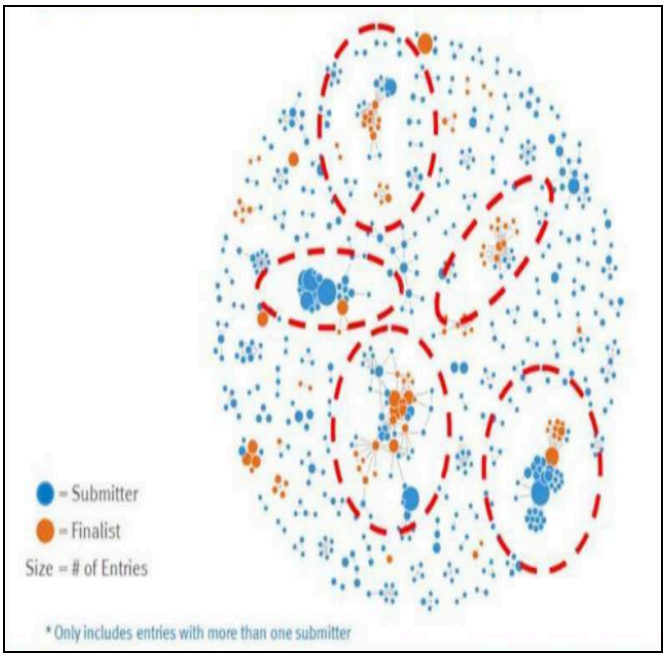

Fig 4.1: Social graph visualization of idea submitters and finalists

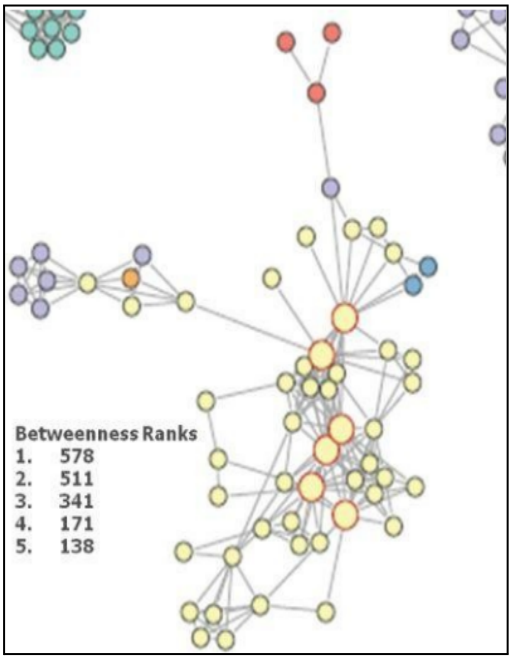

Fig 4.2: Social graph visualization of top innovation influencers

Figure 4.1 illustrates social graphs depicting relationships among idea submitters within GINA, with each color representing innovators from different countries. The large dots with red circles denote hubs, indicating individuals with high connectivity and a high "betweenness" score. Figure 4.2 displays a social graph representing top innovation influencers. The graph exhibits geographic diversity, crucial for validating the hypothesis about geographic boundary spanners. Notably, one individual in the graph stands out with an exceptionally high score compared to others. Upon further investigation, the data scientist discovered significant information about this research scientist's influence within the company, spanning various business units and geographic locations.

Specifically, the research scientist's activities included attending prestigious conferences like ACM SIGMOD and SDM 2012, visiting EMC business units in France, presenting insights from conferences at virtual brown bag sessions attended by employees from multiple countries, and engaging with innovators and researchers at EMC federated companies like Pivotal and VMware. This finding supported the initial hypothesis, demonstrating the ability of data to identify innovators spanning different geographies and business units.

For data visualization and exploration, the team utilized Tableau software, while the Pivotal Greenplum database served as the primary data repository and analytics engine.

Phase 5: Communicate Results

Phase 5 marked the culmination of the GINA project, where the team effectively communicated the results of their analysis and identified impactful findings. The project successfully identified boundary spanners and hidden innovators, prompting the CTO office to launch longitudinal studies for ongoing data collection to track innovation outcomes over extended periods. GINA facilitated knowledge sharing on innovation across various areas within and outside EMC, fostering additional intellectual property cultivation, generating new research topics, and fostering university collaborations in Data Science and Big Data research fields. Remarkably, the project achieved its objectives within a limited budget, leveraging a volunteer force of highly skilled engineers and data scientists.

A key discovery was the disproportionately high density of innovators in Cork, Ireland, identified through the analysis. Despite its relatively smaller size compared to other centers, the Cork COE exhibited remarkable innovation performance, with 15% of finalists and winners hailing from Ireland in a company-wide innovation contest. Further investigation revealed that focused innovation training from an external consultant had significantly enhanced Cork COE's innovation capabilities, resulting in increased idea generation and higher-quality contributions to EMC's innovation efforts. Traditional methods or anecdotal feedback would have likely missed this cluster of innovators, highlighting the efficacy of social network analysis in identifying impactful contributors. These findings were disseminated internally through presentations, conferences, social media, and blogs, emphasizing the importance of data-driven insights in driving innovation and organizational success.

Phase 6: Operationalize

The analytics conducted on a sandbox filled with notes, minutes, and presentations from EMC's innovation activities yielded valuable insights into the company's innovation culture. Key findings from the project include:

1. The CTO office and GINA require more data in the future, necessitating a marketing initiative to encourage people to share their innovation and research activities globally.
2. Sensitivity of some data necessitates consideration of security and privacy concerns regarding model access and result visibility.
3. Apart from model execution, a parallel initiative is needed to enhance basic Business Intelligence activities, such as dashboard development, reporting, and queries on research activities worldwide.
4. Continuous evaluation and retraining of the model post-deployment are essential to maintain its effectiveness.
5. Assessing benefits and defining a process for model retraining are primary objectives in this phase.
6. The project showcased how analytics can offer insights into traditionally challenging-to-measure projects, informing investment decisions in university research projects and identifying hidden, high-value innovators.
7. The CTO office developed tools to aid idea submitters, including topic modeling-based recommender systems to help refine proposals for new intellectual property.
8.

Overall, the project highlighted the transformative potential of analytics in driving innovation and decision-making processes within EMC, underscoring the importance of data-driven insights in shaping organizational strategies and initiatives.

Conclusion:

This project tackled the challenge of measuring and enhancing innovation within companies by evaluating informal social networks to identify boundary spanners and influential individuals within innovation subnetworks. Through the application of advanced analytical methods, the project provided objective, fact-based insights into this seemingly nebulous problem.

Additionally, the project underscored the importance of supplementing analytics with a separate datastore for Business Intelligence reporting, facilitating decision-making and enabling awareness of global discussions and research initiatives across dispersed team members. Moreover, it emphasized the value derived from data and subsequent analysis, highlighting the necessity of formal marketing programs to encourage individuals to share their innovation and research activities globally. Knowledge sharing emerged as a critical factor, enabling the project to uncover hidden innovators within the company and perform comprehensive analyses.

Analytic Plan from the EMC GINA Project

Components of Analytic Plan	**GINA Case Study**
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities.
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis.
Result and Key Findings	1. Identified hidden, high-value innovators and found ways to share their knowledge 2. Informed investment decisions in university research projects 3. Created tools to help submitters improve ideas with idea recommender systems

REFERENCES

1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data: EMC Education Service
2. Data analytics life cycle at the Global Innovation Network and Analysis (GINA):

   https://nishjay711.medium.com/data-analytics-life-cycle-at-the-global-innovation-network-and-analysis-gina-ca2734823234