1-2021

Understanding the inter-domain presence of research topics in the computing discipline

Subhajit DATTA
Singapore Management University, subhajitd@smu.edu.sg

Rumana LAKDAWALA
BITS Pilani

Santonu SARKAR
Birla Institute of Technology and Science (BITS)

IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 9 December 2017; revised 17 July 2018; accepted 24 July 2018.
Date of publication 0 0000; date of current version 0 0000.

Digital Object Identifier 10.1109/TETC.2018.2869556

Understanding the Inter-Domain Presence of Research Topics in the Computing Discipline

SUBHAJIT DATTA, RUMANA LAKDAWALA, AND SANTONU SARKAR

S. Datta is with the School of Information Systems, Singapore Management University, Singapore 178902
R. Lakdawala is with BITS Pilani, Goa 403726, India
S. Sarkar is with Birla Institute of Technology & Science, Pilani K K Birla Goa Campus, Zuarinagar Goa 403726, India
CORRESPONDING AUTHOR: S. DATTA (subhajit.datta@acm.org)

ABSTRACT The very nature of scientific inquiry encourages the flow of ideas across research domains in a discipline. Research topics with higher inter-domain presence tend to attract higher attention at individual and organizational levels. This is more pronounced in a discipline like computing, with its deeply intertwined ideas and strong connections with technology. In this paper, we study corpora of research publications across four domains of the computing discipline – covering more than 150,000 papers, involving more than 200,000 authors over 55 years and 175 publication venues – to examine the influences on inter-domain presence of research topics. We find statistically significant evidence that *higher* collective eminence of researchers publishing on a topic is related to *lower* inter-domain presence of that topic, *fewer* authors publishing on a topic relate to the topic being likely to have *higher* inter-domain presence, while topics belonging to *more* close-knit clusters of topics are likely to have *lower* inter-domain presence. Our results can inform decisions around defining and sustaining research agendas and offer insights on the progression of the computing discipline.

INDEX TERMS Computing, research topics, domains, latent Dirichlet allocation (LDA), statistical models

I. INTRODUCTION AND MOTIVATION

Scientific research disciplines fragment over time [1]. As existing research problems are addressed, newer problems open up, spawning sub-communities of researchers. These "invisible colleges" [2] focus on domains *within* disciplines. However, increasing fragmentation makes it more difficult for individual researchers to remain continuously updated with the latest development in every area of interest within a discipline. Thus, researchers are always facing the *hedgehog versus fox* [3] dilemma; whether to strive to be an expert in a focused field, or aim for familiarity with a wide range of ideas. Addressing this question is central to setting up and sustaining research agendas over the course of individual careers and organizational trajectories. A necessary step in that direction is to examine how research topics across domains overlap, as they relate to the common foundations of the discipline. In this paper we take *computing* as our discipline of interest and study how research topics in the *domains* of *artificial intelligence*, *databases*, *operating systems*, and *software engineering* overlap within the computing discipline. Specifically, we study the following research question:

> *What are the factors that relate to a research topic's high inter-domain presence?*

Understanding these factors have notable implications for individuals and organizations. For young researchers entering a discipline, the map of the research ecosystem often appear imperceptible. A deeper understanding of how research topics across domains connect with one another, and whether and why some topics have higher inter-domain presence than others can be a valuable mechanism for choosing specific research problems. Given the varying half-lives of research topics [4], it is quite natural for the research landscape of a domain to change – often dramatically – within one researcher's active working life. Thus, for veteran and tyro researchers alike, a sense of what leads to a particular topic having large inter-domain presence can be valuable. Academic and industrial research organizations continually need to evaluate proposals and make funding decisions. With the recent thrust in global collaboration, research proposals with a strong inter-domain appeal are often favourably considered [5]. In the evaluation of such proposals, results from our study can offer insights into whether a proposed research project offers sufficient breadth across domains.

A. SCOPE OF THE STUDY

To understand research topics' inter-domain presence, it would be ideal to examine topics across *all* domains of a discipline. However what constitutes "all" is far from a settled question. Even for a relatively young discipline like computing, the intra-discipline domain space has been expanding over time, as evident in the regular initiation of new workshops, working conferences, and tracks at premier conferences, and the launching of journals focussed on specific domains. This is far from being merely a recent phenomenon, as established by Price through his investigation of the long standing scientific disciplines [2]. With this background, we have selected four domains within the computing discipline for our study, as mentioned earlier. Our choice was guided by the following considerations:

- *Artificial intelligence* (AI): In the middle of the 20th century, initial interest in artificial intelligence brought with it great promise of *rapid*, game-changing innovations. However, much of that promise remained unfulfilled in the subsequent decades [6]. In recent times, the potential of autonomous vehicles and other factors have led to a notable resurgence of AI; something that is noticed – and often, feted – even outside the research community [7]. Thus AI embodies a domain with a clearly discernible cycle of waxing, waning, and then renewed waxing research interest. This endows AI with distinct characteristics as a domain within computing.
- *Databases* (DB): How data is curated, processed, persisted, and accessed has changed immensely over the last fifty years. During this time, DB researchers have investigated a broad swath of research questions, starting from the conception of relational and other models of data storage, to the recent investigations around big data [8]. The DB domain subsumes areas such as information retrieval and data mining, which are increasingly attracting research attention in this day and age of easy availability of large scale data. In this context, as "data science" gains traction among researchers and practitioners, how DB relates to other computing domains poses an interesting question.
- *Operating System* (OS): Operating systems research reflects the progression of a vital aspect of the computing discipline, from the days of mainframes to the wave of personal computers, and then to today's ubiquity of hand-held devices. Many fundamental principles of computing have come out of OS research. However, the OS research community has remained close-knit and focussed to a relatively large extent [9]. With this background, we believe our research question can shed interesting light on the inter-domain characteristics of OS research topics.
- *Software engineering* (SE): SE is one of those rare domains – if not the only one – which has undergone a distinct metamorphosis in character over the decades of its existence: from being predominantly theoretical to increasingly empirical, while remaining within the general ambit of the computing discipline [10]. With theoretical and empirical disciplines having distinct research mores [11], SE offers a unique test bed for studying the overlap of its research topics with other domains.

B. CONTRIBUTIONS AND KEY FINDINGS

In this paper, we extract *topics* from corpora of research publications across AI, DB, OS and SE domains from publicly available bibliographic repositories. Our corpora of research publications from these domains consist of 152,510 papers across 216,337 authors covering 175 publication venues across 55 years. To the best of our knowledge, this is the largest corpora analysed for a study of similar scope.

We find evidence of higher collective eminence of researchers publishing in topics, to be related to lower interdisciplinary presence of such topics. This reflects on some key decisions researchers need to make early in their careers. Pursuing research interests that spread across various domains may help a researcher develop a portfolio of diverse results. However, this policy may not necessarily lead to higher scores in established research impact metrics.

Perhaps as a concomitant phenomenon, we find evidence that topics with higher inter-domain presence do indeed relate to fewer researchers publishing in those topics. This may be an indication that researchers subconsciously sense the inverse relation between individual eminence and inter-domain presence of topics that our study reveals, and take actions commensurate with a striving for higher eminence. Additionally, it can also point to the rarity of qualities – possessed by only a few researchers – that successful conduct of interdisciplinary research calls for.

Additionally, we find that topics which are positioned in close contextual proximity with other topics are more likely to have lower inter-domain presence. This can inform the shaping of research agendas at various points in a research career; whether a researcher turns out to be a hedgehog or a fox [3], appears to be closely tied to the type of research topic(s) she chooses to pursue.

C. UTILITY

Different aspects of the study reported in this paper offer a variety of benefits to researchers and practitioners. In the next section, we present a methodology for extracting research topics from corpora of publications in different domains and then quantifying the extent of inter-domain presence of the topics. Given the nature of our study design, it is expected that any such methodology will encapsulate various choices of consequence. By discussing the development of our methodology in detail – including the considerations behind the choices – we present perspectives which will facilitate other researchers to adopt or adapt our methodology. In the subsequent section on results and discussion, statistical models are developed and analysed to address the research question. The identification and computation of model variables offer different points of view on the factors influencing inter-domain presence of topics. These factors

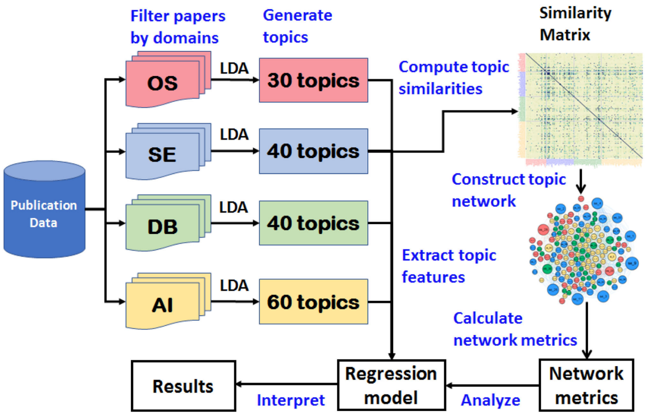


FIGURE 1. Outline of the methodology.

can inform many of the tactical as well as strategic choices researchers need to make in their careers. Finally, the discussion of our results, as well as highlighting the threats to their validity offers a balanced evaluation of the conclusions from the study. This can serve as a foundation for similar studies in other disciplines.

The paper has been organized in the following way: In Section II we describe a methodology based on latent Dirichlet allocation to extract the topics and other relevant information. Next, we propose the notion of topic similarity; and based on it, create a network of topics across the aforementioned four domains. In Section III, we build statistical models to identify parameter(s) that influence the inter-domain presence of a topic. In Section IV we identify the threats to the validity of our results. Related work is outlined next in Section V, and the paper ends with summary and conclusions in Section VI.

II. METHODOLOGY

In this section we describe the methodology of this study; Figure 1 outlines the key components of our approach. As described in Section II-A, we first access publicly available scientific publication data in the computing discipline, and then segregate the data into the domains of our interest. Next, sets of topics are extracted for each domain using the procedures discussed in Section II-B. To specify the extent of inter-domain presence of topics, a quantitative notion of *topic similarity* is developed in Section II-C, where we further evaluate different approaches and establish the suitability of the the specific approach chosen for this study. From topic similarity, we construct a network of topics using the protocol specified in Section II-E. Subsequently, we highlight the calculation of network metrics, and development of a regression model in Sections II-F and II-G, respectively.

A. DATA AND DOMAINS

As mentioned earlier, we consider the computing domains of artificial intelligence, database, operating systems, and software engineering in this study. We collected information about research papers – including their abstracts – published in prominent venues of these domains from sources such as *Microsoft Academic Graph*[1] and *AMiner*.[2] Thus, the corpus for our study $\mathcal{D} = \bigcup \mathcal{D}_i$ comprises of papers, authors, publication venues and abstracts from these four domains (we denote the corpus pertaining to a domain i by $\mathcal{D}_i$). We create an initial vocabulary set from the words obtained from the title and abstract of each paper. Next, this vocabulary set is pruned by stemming each word to its root form, and by identifying and removing frequently used terms as stop words. This pre-processing step is essential for the subsequent analysis as it eliminates words conveying little semantic content, and semantically related words are aliased under the same root if they share the same canonical form [12]. After performing this standard pruning process, we filter the resultant keywords further, based on their term and the document frequency values.

Identifying the key ideas or a research topic on which a paper has been published is a non-trivial task. Additionally, there are no established approaches for identifying the domain to which a research topic belongs, and quantifying the extent to which research ideas in a topic have come from different domains. To address these issues we have to (i) define and identify research topics from publication corpora, (ii) devise a method to decide on the domain (or domains) to which a topic belongs. (iii) devise a method to quantify the inter-domain presence of a topic. We outline these steps in the next subsections.

B. TOPIC DISCOVERY AND NAMING

As in our previous work in this area [4], as a proxy for research ideas, we consider automatically discovered *topics* from our data-set using latent Dirichlet allocation (LDA). LDA has been widely used in various applications to extract topics from large corpora of text documents [13]. Briefly, LDA considers a document to be a mixture of topics $T = \{\tau_1 \ldots \tau_K\}$ and each topic is characterized by a distribution over terms. From a corpus $\mathcal{D}$, LDA outputs ϕ = multinomial distribution over terms for topics and θ = multinomial distribution over topics for documents.

The effectiveness of LDA to segregate document collections into relevant themes has been demonstrated when the number of topics is known a priori [14]. However the difficulty arises when the number of topics is not known. *Perplexity* is a commonly used measure to evaluate how well a statistical model describes a data set, with lower perplexity denoting a better fit to the probabilistic model. We have used the perplexity based measure similar to the approach described in [13] where the authors have used perplexity to compare the relative strengths of several topic models.

After identifying topics and their associated keywords, it is important to able to get an intuitive sense of what each topic represents. Thus, ascribing a meaningful label to each topic can be helpful. Towards that end, we employ a heuristic-based

[1]https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[2]https://aminer.org/

approach to label topics generated from LDA for each of the four domains. While creating the bag of keywords from a corpus, we not only include the individual terms (1-gram), but we also create a set of 3-grams from these keywords based on their relative positions in documents. Once this bag of words (a word can be a 1-gram or a 3-gram) is processed by LDA to generate a topic model, for each topic, we select the top 3-gram term as the label of the topic from the probability distribution of terms over that topic. The labels are then evaluated manually to check for coherence between labels and associated keywords for each topic. During manual inspection, some of the topic labels are modified by considering the keywords present in the corresponding topic.

C. TOPICS ACROSS DOMAINS

We considered two approaches to identify the domain a topic belongs to. In the first approach we maintain a single corpus with all the papers from the four domains. In the second approach, we consider separate corpus $\mathcal{D}_i$ for a particular domain i. In the following discussion, we will refer to approach with the unified corpus as the *combined* approach and the second approach as the *partitioned* approach. The *partitioned* approach, in a sense, honors the natural boundaries created across different domains; where domains start independently and then cross-pollination of ideas across domains takes place over time. Since the publication venue of a paper belongs to one particular domain, the association between a published paper and the domain is always preserved, irrespective of approaches. In the *combined* approach, the final, pruned vocabulary set V comprises of 60,000 terms whereas in the case of the *partitioned* approach, we pruned the text corpus for each domain $\mathcal{D}_i$ to retain roughly the top 20,000 terms in the vocabulary V^i for each domain i. In the *partitioned* approach, since each domain has specific, broad focus, it makes sense to eliminate words that have less discriminatory power. For instance, in SE, a word like "software" has little specific significance, whereas the word "database" is relatively more important. On the contrary, in the DB domain, the word "database" will not be very discerning. In view of this, we have modified each domain specific stopword list to include such words.

The number of topics for each domain was determined by evaluating the perplexity of the corresponding topic models iteratively. We finally arrived at 60 topics for AI, 40 topics for DB, 30 topics for OS, and 40 topics for SE; this gives us 170 topics in total across the four domains. We have kept 170 topics for the *combined* approach as well, for the ease of comparative analysis.

D. DETERMINING INTER-DOMAIN PRESENCE OF TOPICS

The position of topics in a knowledge space is obtained from the distribution of terms characterizing each topic, which can be represented as a vector in the vocabulary space. The idea behind representing a topic as a vector in high dimensional space of terms is to investigate the similarity between the ideas represented by these topics. In the case of the *combined* approach, there is only one vocabulary set V; thus the vector model is trivial

$$\vec{v'}_\tau = [w'_{1,\tau}, w'_{2,\tau}, \ldots, w'_{M,\tau}],$$

where M is the total number of terms in V and $w'_{i,\tau} = \phi_{\tau,i}$ for the ith term in V.

The vector model $\vec{v}$ for the *partitioned* approach is similar to the earlier one with a subtle difference, where we need to consider four domain vocabulary sets instead of one. Let N be total number of terms present in the combined space of the four domain's vocabularies $V^{AI} \cup V^{SE} \cup V^{OS} \cup V^{DB}$. Each dimension of the topic vector $\vec{v}_\tau$ corresponds to a term in the combined vocabulary and the term weights corresponds to the probability of that term in topic τ (of a particular domain i) only if the term is included in the vocabulary of that domain. Thus: $\vec{v}_\tau = [w_{1,\tau}, w_{2,\tau}, \ldots, w_{N,\tau}]$ where

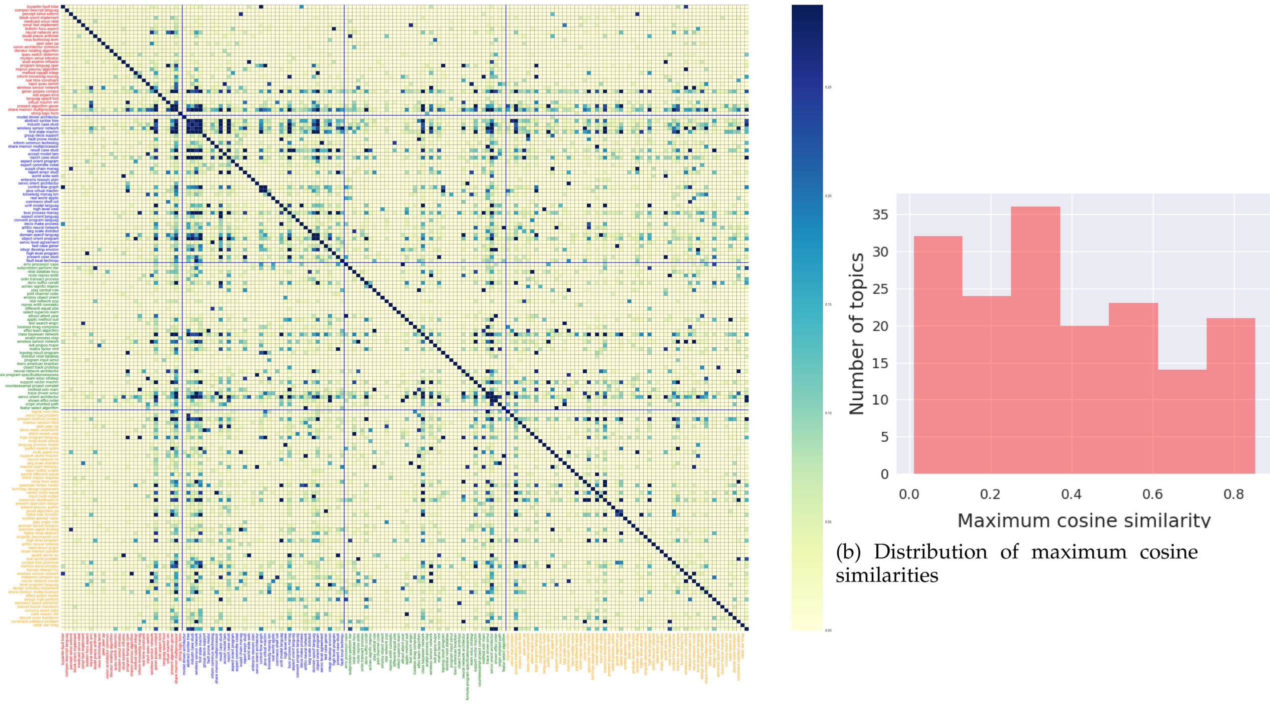
$$w_{i,\tau} = \begin{cases} \phi_{\tau,i}, & \text{if } term_i \in V^i \\ 0, & \text{otherwise.} \end{cases}$$

From a corpus $\mathcal{D}$, LDA generates ϕ = multinomial distribution over terms for topics and θ = multinomial distribution over topics for documents. Thus $\phi_{\tau,i}$ is the probability of term i belonging to topic τ. To evaluate the similarity between ideas represented by two topics, *cosine similarity* can be calculated between vectors of the topics. Cosine similarity, given by

$$\cos(\vec{v}_{\tau a}, \vec{v}_{\tau b}) = \frac{\vec{v}_{\tau a} \cdot \vec{v}_{\tau b}}{\| \vec{v}_{\tau a} \| \cdot \| \vec{v}_{\tau b} \|},$$

indicates the angle between two topic vectors $\vec{v}_{\tau a}$ and $\vec{v}_{\tau b}$, measures of how "similar" they are, which in turn, reflects on the extent of congruence between their terms. Evidently, the cosine similarity values will be in the range $[0 \cdots 1]$. The reasons behind using cosine similarity over other similar indicators include the efficiency of calculating it over high dimensional sparse vectors of topics in our corpora, and the fact that it is a robust metric, typically used in comparing text-based vectors.

In order to compare the efficacy of the *combined* and the *partitioned* approaches, we compute the pairwise cosine similarity across all pairs of the 170 vectors for both the approaches and generate 170×170 *topic similarity matrices* S' and S for the *combined* and *partitioned* approaches respectively. Each cell of the matrix corresponds to the cosine similarity between the topics corresponding to the row and column of that cell, that is: $S'[i,j] = \cos(\vec{v'}_{\tau i}, \vec{v'}_{\tau j})$ for the *combined* and $S[i,j] = \cos(\vec{v}_{\tau i}, \vec{v}_{\tau j})$ for the *partitioned* approach. Figure 2(a) represents S *A heat map*, with higher cosine similarity values marked with darker shades. The diagonal of the map obviously is in the darkest shade, since $S[i,i]$ for all i will always be 1. Figure 2(a) shows many dark shades indicating that there are several topic pairs which are close to each other. In comparison, we noticed relatively far



(b) Distribution of maximum cosine similarities

(a) Heat-map of the topic similarity matrix, S

FIGURE 2. Cosine similarity analysis.

fewer dark cells other than on the trivial diagonal line, in the corresponding heat map of the combined approach.

A suitably tuned LDA based approach strives to identify the latent topics which are not diffused, where the keyword set gets partitioned into a set of topics with minimal overlap between two topics; the nature of the corpus obviously having a bearing on the extent of overlap. Let us now examine the ramifications of using the *combined* approach for topic extraction vis-a-vis the *partitioned* approach in the context of our study, from the following perspectives:

1. When we mix papers from all four domains and create a topic model on the combined set of papers – where there are several papers from different domains sharing the same set of keywords – the LDA algorithm will most likely extract a topic (among other topics) such that papers from different domains will be associated with that topic due to the commonality of keywords. Furthermore, since LDA aims to segregate the keyword space, once we get a set of topics from the combined set of papers, we can expect to get minimal commonality among topics. We can observe this in the heat map of the *combined* approach in Figure 2(b), with its significantly sparse instances of darker shades, as compared to Figure 2(a). *This confirms the fact that the inter-topic similarity between topics obtained through the combined approach is indeed very low in comparison with the partitioned approach*.
2. To further understand the distribution of the cosine similarity values, we compute the maximum cosine similarity values of each topic with respect to the others, as *max cossim*$(\tau) = \max\{\cos(\vec{v}_\tau, \vec{v}_{\tau_i}) | \forall \tau_i \neq \tau\}$. Thus, we get 170 maximum cosine similarity values each for the topics obtained from the *combined* and *partitioned* approaches. While comparing the frequency distributions of the maximum cosine similarity values from the partitioned and combined approaches, we noticed the the latter is highly skewed towards the right, whereas the former is relatively closer to a uniform distribution. In the combined approach, more than 100 topics have at most 0.01 cosine similarity with any other topic. However, in the partitioned approach, significant number of topics have maximum similarity over 0.4 and some topics have maximum similarity as high as 0.8 with other topics. Figure 2(b) shows the frequency distribution of the maximum cosine similarity values of the partitioned approach. *This comparision clearly indicate that the partitioned approach offers more meaningful results that the combined approach, when cosine similarity is used to measure the extent of interdomainness of topics.*

In light of the above discussion, we selected the *partitioned* approach for this study.

E. CONSTRUCTING AN INTER-DOMAIN TOPIC NETWORK

To further investigate the implications of topic similarities, we constructed a *topic network* (*NW*) using the following method: The vertices of *NW* are the 170 topics across domains. There exists an edge between topics τ_i & τ_j of *NW* if the similarity between them is in the upper *quartile* (Q_3) – that is, the top 25 percent – of topic similarities in the

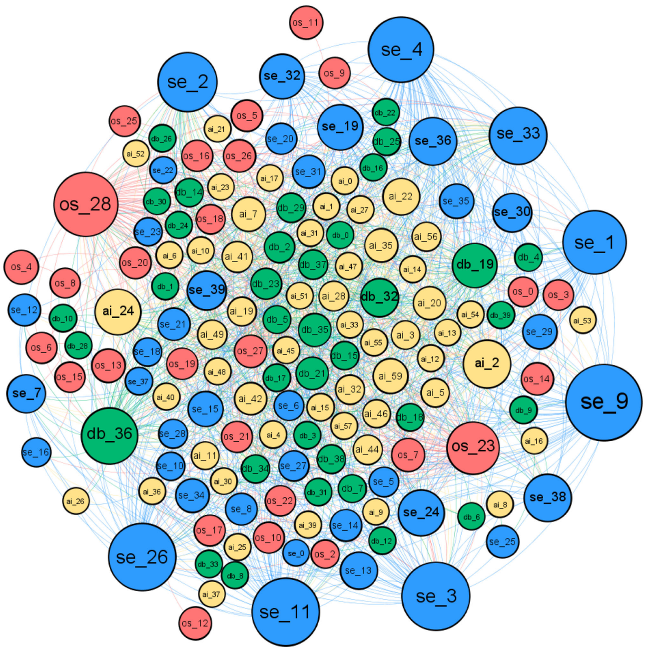


FIGURE 3. **Inter-domain topic network: Vertices represent topics and are color coded by the domains. The vertex identifiers indicate the domains and topic numbers. Vertices are sized by the corresponding topic's *interdomainness*.**

similarity matrix. $Q_3(S^{ut})$ indicates the upper quartile of S^{ut}, which is the upper triangular matrix of S, since topic similarity is a symmetric entity. Using the upper quartile as a differentiator of high-value groups is widely used in diverse fields. For example, an individual's mean skinfold thickness in the upper quartile of a cohort has been taken to be an indicator of obesity [15]; trustworthiness of e-commerce vendors has been decided on the basis of whether they are in the upper or lower quartile on a scale of relevant measurements [16].

This set of edges, E of NW can be formally defined as

$$(i,j) \in E \quad if \ S_{i,j} > Q_3(S^{ut}) \ \wedge \ i,j \in NW.$$

Figure 3 shows a particular visualization of NW, relevent to this study.

F. CALCULATING NETWORK METRICS

Constructing the topic network enables us to evaluate various aspects of the putative "network effects" as they relate to the context of our study; many of these effects have been captured in established metrics in the network science literature [17]. As discussed in detail in Section III-B, we consider some such metrics as we develop the statistical model to address our research question.

G. DEVELOPING A REGRESSION MODEL

To examine a research question in the light of empirical evidence, statistical model(s) needs to be developed. In this study, we use a regression model to study the influences of various factors on the inter-domain presence of topics. In Section III-C, we weigh the considerations in the choice of the modeling paradigm, followed by presentation and discussion of results from the model.

H. SUMMARY OF THE STUDY DESIGN

The methodology outlined in the preceding sections and further developed and applied in the next section can be summarized as:

- Extract *topics* using LDA, from corpora of research publications in the AI, DB, OS, SE domains.
- Construct a *similarity matrix* by computing cosine similarity between all topics.
- Construct a *topic network* from the similarity matrix.
- Define a measure of a topic's inter-domain presence.
- Identify characteristics that relate to a topic's inter-domain presence.
- Develop a statistical model to understand these relationships.

III. RESULTS AND DISCUSSION

A. OBSERVATIONS FROM SIMILARITY MATRIX AND TOPIC NETWORK

Having defined and constructed the topic similarity matrix S and the topic network NW, let us observe their characteristics. With reference to Figure 2(a), we note that S is symmetric around its main diagonal. Focusing our attention on the upper triangle, we see several dark regions of varying intensity dispersed within and across the domains. It is not unexpected that topics *within* a domain would share some degree of similarity; that is precisely why they collectively constitute a domain. However, regions of darkness corresponding to a particular topic *outside* its own domain are interesting, as they indicate varying degrees of that topic's inter-domain presence. Table 1 provides a list of topic-pairs that have relatively high inter-domain presence, with correspondingly high cosine similarities. Some of the interesting observations are:

- A topic pertaining to software engineering domain seems to be close to world wide web. As mentioned earlier, we considered the suggestions from human experts to override the 3-gram based labelling for this topic. Topics with name "service orient architectur" are also present in multiple domains.
- The topic with the label "wireless sensor network" cuts across multiple domains and plays a significant role in the interdomainness.
- Topics related to formal methods e.g., finite state machine, have similarities with topics from other domains.

In addition, we show few intra-domain topics that have high cosine similarity values – in the upper quartile – in Table 2.

To capture the characteristic of interdomain presence of topics, we define the *interdomainness* of a topic to be the *median of the cosine similarities of that topic with all other topics across domains*. Given the varied distributions of the cosine similarities of the topics, the median – rather than the mean – was selected to be a more effective measure of central tendency.

TABLE 1. Example of inter-domain topic similarity.

Label τa	Label τb	Keywords τa	Keywords τb
studi examin influenc (OS)	world wide web (SE)	servic-internet-attack-encod-multicast-provid-script-messag-distribut-traffic-inform	servic-provid-intellig-fault toler-middlewar-servic provid-design-taxonomi-web
servic orient architectur (SE)	formula program specifications (DB)	architectur-enterpris-layer-style-servic orient-architectur design-level-crosscut	architectur-unit-behavior-treat-design-axiomat-crosscut-hoar-tecton
finit state machin (SE)	class Bayesian network (DB)	algorithm-method-problem-approach-gener-propos-techniqu-function-state-result	method-algorithm-model-approach-propos-Bayesian-gener-techniqu-cluster-result
test case gener (SE)	genet algorithm ga (AI)	test-case-test case-gener-coverag-techniqu-suit-test suit-execut-effect	test-discuss-descript-investig-genet-algorithm-oper-evolutionari-present-genet algorithm
wireless sensor network (OS)	wireless sensor network (DB)	network-sensor-node-wireless-sensor network-protocol-commun-wireless sensor-wireless sensor network-energi	network-problem-sensor-object-node-address-challeng-locat-wireless-commun
bulletin focu aspect (OS)	high level program (SE)	secur-java-polici-aspect-track-privat-overflow-focu-pointer-avail	secur-question-attack-trust-answer-reengin-threat-light-offic-answer question
servic orient architectur (DB)	share memori multiprocessor (OS)	model-perform-manag-process-servic-develop-architectur-requir-support-tool	perform-design-oper-time-comput-process-memori-inform-distribut-program
present algorithm gener (OS)	finit state machin (SE)	algorithm-problem-implement-present-method-linear-comput-function-solv-block	algorithm-method-problem-approach-gener-propos-techniqu-function-state-result
high level view (SE)	languag process model (AI)	rang-trade-tabl-wide-wide rang-percent-profession-simplic-lose-circumst	rang-wide-aris-scan-wide rang-acquisit-bank-fortran-broad-quadrat
program languag oper (OS)	logic program languag(AI)	languag-program-program languag-graph-untrust server-architectur-garbag-cube-edg-support	program-languag-knowledg-logic-reason-graph-semant-formal-represent-theori
wireless sensor network (SE)	share memori multiprocessor (OS)	time-network-distribut-perform-simul-comput-parallel-real-real time-commun	perform-design-oper-time-comput-process-memori-inform-distribut-program
finit state machin (SE)	gener purpos comput (OS)	algorithm-method-problem-approach-gener-propos-techniqu-function-state-result	model-file-gener-method-present-queri-distribut-algorithm-form-approach
wireless sensor network (AI)	wireless sensor network (DB)	commun-network-sensor-distribut-fault-node-protocol-wireless-messag-flexibl	network-problem-sensor-object-node-address-challeng-locat-wireless-commun
servic orient architectur (DB)	gener purpos comput (OS)	model-perform-manag-process-servic-develop-architectur-requir-support-tool	model-file-gener-method-present-queri-distribut-algorithm-form-approach
servic orient architectur (DB)	servic orient architectur (SE)	model-perform-manag-process-servic-develop-architectur-requir-support-tool	architectur-enterpris-layer-style-servic orient-architectur design-level-crosscut

In Figure 3, the vertices of the network are sized according to the corresponding topic's *interdomainness*; it is evident there is a wide range of variation in the topics' *interdomainness*. Thus, to address the research question introduced in Section I, we need to identify a set of topic characteristics that can help us explain the variance in the topics' *interdomainness*. With this objective, we identify the following factors, on the basis of our general perception of the study setting, as well as results from existing studies [4],[18]:

1) The *eminence* of researchers who publish on a topic.
2) The number of *papers* published on that topic.
3) The number of *venues* where papers on that topic are published.
4) The number of *authors* publishing papers on that topic.
5) The *domain* which the topic belongs to.
6) The extent to which the topic comes *in between* diverse other topics.
7) The extent to which the topic belongs to close-knit *clusters* of other topics.
8) The *age* of the topic in terms of the number of years papers are being published on that topic.

TABLE 2. Example of intra-domain topic similarity.

Label τa	Label τb
particl swarm optim (AI)	markov decis process (AI)
result case studi (SE)	report case studi (SE)
gener purpos comput (OS)	share memori multiprocessor (OS)
wireless sensor network (AI)	neural network model (AI)
abstract syntax tree (SE)	finit state machin (SE)
close form solut (AI)	constraint satisfact problem (AI)
world wide web (SE)	servic orient architectur (SE)

B. COMPUTING MODEL VARIABLES

For a quantitative analysis of the influences on topic *interdomainness* using statistical models, we need to identify suitable

TABLE 3. Pearson correlations coefficients of model variables with the dependent variant - *interdomainness*.

Variable	Correlation
Eminence	−0.328
Papers	0.292
Venues	0.168
Authors	0.122
Betweenness	0.739
Clustering	−0.369
Age	−0.249

TABLE 4. Descriptive statistics of the model variable.

Variable	Mean	Standard deviation	Median
Interdomainness	0.01	0.016	0.004
Eminence	10.95	5.666	10
Papers	5586.353	5064.565	4147
Venues	37.935	11.2	36
Authors	1.100×10^4	9503.367	8503.5
Betweenness	0.005	0.01	0.001
Clustering	0.629	0.196	0.656
Age	43.547	8.389	41

proxies for each of the above factors, which can be computed from our corpora or the topic network. Towards that end, we define the following *mapping* between a topic and paper:

Given a domain D, a topic τ_k extracted from D, and a paper p in D, p *belongs* to τ_k if the paper-topic probability θ^D_{p,τ_k} is in the upper (that is, top 25 percent) quartile of all paper-topic probabilities of D. The arguments for using the upper quartile to represent high values in diverse research settings have been outlined earlier in Section II.

Thus, using the above mapping between papers and topics, with the threshold, $\gamma = Q_3(\theta)$, for each topic τ_k, we can generate a set of papers, P^{τ_k} that belong to τ_k, such that

$$P^{\tau_k} = \{p | p \in P^D,\ \theta_{p,\tau_k} > \gamma\}.$$

Given the set P^{τ_k}, we compute the following model variables as proxies for our factors of interest identified above.

1) *Eminence:* The collective eminence of authors publishing papers on a topic is calculated as the median of their *h-indices* [19]. The h-index is selected as it is a well-established measure of researcher impact; we discuss the implications of this choice in Section IV.
2) *Papers:* The number of papers that belong to a particular topic is represented by this variable.
3) *Venues:* For this variable, we count the number of unique venues in which papers belonging to a topic are published.
4) *Authors:* We count the number of unique authors publishing papers on a topic to represent this variable.
5) *Domain:* As mentioned earlier, our study includes four domains - AI, DB, OS, SE. We use three "dummy variables" $D.x$, $D.y$, and $D.z$ to capture effects that are specific to the domains, $n - 1$ dummy variables being sufficient to model the effects of n categorical variables [20].
6) *Betweenness:* The notion of "betweenness" reflects on how important a particular vertex is, as an intermediary between other vertices in a network. It is measured by the *betweenness centrality* of a vertex, which is the proportion of geodesic paths between all pairs of vertices in the network, which includes that vertex [21]. In the context of this study, *betweenness* of a topic is computed as the betweenness centrality of the corresponding vertex in the topic network.
7) *Clustering:* The clustering coefficient (C_v) for a vertex v in a network is defined as the ratio of the actual number of edges existing between its neighbors and the maximum number of such edges that can exist [21]. Thus, if v has a degree of k_v, that is, there are k_v vertices directly connected to v, the *maximum* number of edges between these k_v vertices is k_v *choose* 2 or $k_v * (k_v - 1)/2$. If the *actual* number of such edges existing is N_v, then $C_v = 2 * N_v / k_v * (k_v - 1)$. Evidently, the clustering coefficient indicates how much a particular vertex is included in clusters within the network. In our study setting, the clustering coefficient of a topic reflects on the topic's embeddedness in community structures in the topic network.
8) *Age:* We calculate the age of a topic as the elapsed time in years between the first publication and last publication in that topic, within our study period.

C. CHOICE OF MODELLING PARADIGM

In Table 3 we present the correlations between *Interdomainness* and the other variables identified above. The descriptive statistics of these variables are given in Table 4. To understand how different factors *collectively* relate to the inter-domain presence of topics, we develop a linear regression model with *Interdomainness* as the dependent variable, and the others as independent variables.

We initially considered developing a Poisson regression model. In a Poisson distribution, the mean equals the variance, which is the single parameter defining the distribution. Over-dispersion – violation of the strong assumption of the equality of variance and mean – is a major threat to the validity of Poisson regression [22]. As this is present in our study, we chose multiple linear regression as the modelling paradigm.

Multiple linear regression rests on the assumptions of linearity, normality, and homoscedasticity of the residuals, and absence of multicollinearity between the independent variables. The residual properties can be verified using the histogram, Q-Q plot and scatter plot of the standardized residuals. We found the variance inflation factors (VIF) for the multiple linear regression model variables to be within permissible limits, thus indicating that multicollinearity does not invalidate our model.

Table 5 presents the model parameters. As specified in the table caption, the signs in the "sig level" column signs denote ranges of their respective p values. The p value for each

TABLE 5. Modelling the influences on inter-domain presence of topics.

	Coefficient	Sth error	Sig level
Intercept	0.017	0.056	-
Eminence	-0.004	0.002	**
Papers	1.783×10^{-5}	2.389×10^{-6}	****
Venues	0.001	4.871×10^{-4}	**
Authors	-7.33×10^{-6}	1.311×10^{-6}	****
D.x	0.081	0.023	****
D.y	0.113	0.016	****
D.z	0.066	0.025	***
Betweenness	1.791	0.224	****
Clustering	-0.044	0.01	****
Age	2.429×10^{-4}	0.001	-
		Model parameters	
N		170	
R^2		0.873	
df		159	
F		108.94	
Sig level		****	

*Note: Significance levels "****", "***", "**", "*", and "-", denote corresponding p-value ≤ 0.001, ≤ 0.01, ≤ 0.05, ≤ 0.1, and ≥ 0.1, respectively.*

coefficient is derived from the t-statistic – the ratio of each coefficient and its standard error – and the Student's *t-distribution*. In the table's lower portion, details of the overall model are given: *N* is the number of data points used in building the model – in our case, the total number of topics across domains. R^2 is the coefficient of determination – the ratio of the regression sum of squares and the total sum of squares; an indicator of the goodness-of-fit of a regression model in terms of the proportion of variability in the data that is explained by the model. *df* denotes the degrees of freedom. *F* is the Fisher *F-statistic* – the ratio of the variance in the data explained by the linear model divided by the variance unexplained by the model. The *p* value is computed using the *F-statistic* and the *F-distribution*, and it points to the overall statistical significance of the model. For the coefficients as well as the overall regression, if $p \leq$ *level of significance* (usually taken as 0.05), we conclude that the corresponding result is statistically significant, on the basis of null hypothesis significance testing.

D. DISCUSSION

Let us now discuss the implications of our results. At the outset, we note that our statistical analyses establish correlation. As this is an observational study rather than a controlled experiment, correlation does not necessarily imply causation. However, in our particular study setting, controlled experiments are almost impossible to conduct, as there is no easy way to segregate research topics into control and treatment groups and observe how *interdomainness* differs between groups. Thus, even as we cannot infer causality, the study setting allows us to examine factors that influence a topic's interdomainness and derive useful insights.

With reference to Table 5 we note that the overall regression model is statistically significant ($p \leq 0.05$), and it is able to explain around 87 percent ($R^2 = 0.873$) of the variance of the underlying data. Let us now observe the relationships between each of the independent variables and the dependent variable and discuss their implications.

1) *INTERDOMAINNESS* AND *EMINENCE*

We see that *Interdomainness* has an inverse relationship with *Eminence*, and the relationship is statistically significant. Thus topics which have higher inter-domain presence are more likely to have a pool of authors whose median h-index is relatively lower. This relationship offers interesting interpretations. Price has pointed out that longevity of a researcher is often a reliable proxy for the quanta of his contribution [2]. As the h-index is a cumulative measure, researchers who have been active in research for a longer period are better positioned to acquire higher h-indices. On the other hand, researchers who are just starting out have lower h-indices. The latter also represents a group which is more inclined to explore different domains as they develop their research agendas. Over time, these agendas usually get restricted to few focus areas, in which each researcher strives to be an expert. So, the inverse relationship between *Eminence* and *Interdomainness* may signal that topics that have a higher spread across domains may be the ones that have attracted a relatively younger pool of researchers.

2) PAPERS, VENUES, AND AUTHORS

We observe in Table 5 that higher number of papers and higher number of venues relate to higher levels of *Interdomainness* of topics, and both these relations are statistically significant. These associations are expected, as more expansive reaches of papers and venues can definitely contribute to higher levels of inter-domain presence of a topic.

However, we also see statistically significant evidence that higher number of authors relate to *lower* levels of *Interdomainness*. This seems to contradict the conventional wisdom that a larger number of authors represent a wider variety of research interests, which in turn can lead to a topic stretching across a wider swath of domains. The inverse relationship observed between the number of authors and *Interdomainness* may be interpreted as an indication that relatively fewer researchers concern themselves with inter-domain research, while the majority are focused on specialization!

3) DOMAIN EFFECTS

We also observe that all the dummy variables D_x, D_y, and D_z, representing the differences between the domains, also have a statistically significant, and direct relationship with *Interdomainness* (Table 5). Thus, whether a topic will have higher inter-domain presence is related to the topic's domain. We can derive a sense of this relationship by observing the vertex sizes and colors in Figure 3. It is clear that topics of certain domains are more likely to have higher *Interdomainness*.

4) TOPIC NETWORK PARAMETERS

We see that both *Betweenness* and *Clustering* of topics have statistically significant relationships with *Interdomainness*

(Table 5). However, the relationship with *Betweenness* is direct, while that with *Clustering* is inverse. Topics with higher *Betweenness* are the ones who predominantly act as bridges between other topics in the topic network. Such intermediary positions signify that topics with higher *Betweenness* have traits that allow them to connect topics which are otherwise disparate. This is closely aligned with the spirit of inter-domain presence of topics, which we have sought to capture in the notion of *Interdomainness*. Thus the relation we find between *Betweenness* and *Interdomainness* matches what is expected. *Clustering* is an indication of triadic closure [23]. Topics with higher *Clustering* are deeply enmeshed with similar other topics and are thus more likely to have a smaller inter-domain presence. This is congruent with the relation between *Clustering* and *Interdomainness* we find from our model.

5) EFFECT OF TOPIC AGE

We find that *Age* has a direct relationship with *Interdomainness*; topics which have been around longer are related with higher inter-domain presence. However, since this relation is not statistically significant, we can not interpret its implications.

6) PREDICTIVE POWER OF THE MODEL

Our model also has notable predictive power, as evidenced by the Pearson correlation coefficient of 0.91 between the actual and predicted values of *Interdomainness*. So, given the values of the independent variables for a particular topic, our model can predict its *Interdomainness* with reasonably high accuracy.

IV. THREATS TO VALIDITY

In any empirical study using statistical techniques, *validity* reflects the extent to which the results relate to the conclusions, in terms of certain established criteria [24]. Identifying and discussing threats to validity is thus a key component in understanding a study's usefulness [20]. In this section, we identify and address the threats to *construct validity*, *internal validity, external validity*, and *reliability*.

Threats to *construct validity* relate to concerns arising from the correct measurement of variables. In our study, all model variables are calculated from data available in the public domain, and using established metrics. As described in Section III, we have chosen the median cosine similarity as a proxy for the inter-domain presence of a topic. We recognize that *Interdomainness* of research topics can be measured in other ways using additional information about research ecosystems of the domains we have studied; and other metrics may lead to different results. Similarly, in this study, the collective eminence of the researchers publishing on a particular topic is measured by the authors' median h-index. We recognize that there is no universally accepted metric to measure impact of individual researchers [25]. Our choice of the h-index is informed by its extensive use in recent times [26]. For both *Interdomainness* and *Eminence*, use of the median - instead of the mean - allows us to get an accurate measure of the central tendency, irrespective of the shape of the underlying distribution. So, while a different metric choice for either or both these variables may lead to different results, they do not represent a threat to the current results. As discussed earlier, we have considered two possible approaches, namely a *combined* approach and a *partitioned* approach, while extracting topics using the LDA model and demonstrated that the latter is more suited to the goal of this study.

Threats to *internal validity* arise from a study's systematic errors and biases. As described in Section II, all our variables are calculated using information from curated, publicly available repositories. Thus, common threats to internal validity such as mortality (subjects being removed from a study during the study period) and maturation (subjects changing character during the study outside research purview) are not present in our case. However, our definition of the AI, DB, OS, SE domains by the research publications from particular sets of venues can be a source of bias. Although we believe our corpora covers a large majority of papers from each of these domains, we can not claim to have captured *all* such papers. Given the fact that all of these domains are within the computing discipline, whether a venue exclusively belongs to a particular domain is a matter of judgment. Thus some papers which belong to one of the domains may have been inadvertently left out, while papers from some other domain(s) may have been included. However, we believe such inclusion/exclusion represents a tiny fraction of our corpora of 150,000+ papers and thus does not pose a significant threat to internal validity

Threats to *external validity* come from the extent to which a study's results can be generalized. As discussed in Section I, each of the four computing domains included in this study has a distinct character. Thus we believe our corpora represent the diversity of the computing discipline in a notable way. However, computing does not only include these four domains. The inclusion of other domains in our study can thus lead to different results. So, our results are not generalizable across the entire computing discipline as yet.

Reliability of a study relates to the reproducibility of the results. Given access to the original data source, our results can be readily reproduced.[3]

In our plans for future work, we seek to include additional domains in our study. We also plan to design studies to further investigate some of the interesting relationships between *Interdomainness* and other variables, as indicated by our statistical models.

V. RELATED WORK

Research ideas seldom remain confined within a given discipline. New ideas usually germinate from an existing body of work due to influences from other areas. While this is known and practised by the researchers, to the best of our knowledge, there has been no data-driven study so far, to characterize such influence. However, there has been substantial work to

[3]The computing resources used in this study can be found at: https://github.com/santonus/bigscholarlydata

characterize the growth of publications in a given discipline, the importance of publication venues, impact of published work on subsequent publications based on citation data analysis, and collaboration among authors. In this section, we give an overview of some aspects of that body of work as they relate to the interaction among topics across disciplines.

A. OUTLINE OF EXISTING STUDIES

While the information about publications (title, venue, year of publication, etc.), citations, authors are available as concrete facts, the notion of a research idea to which a published paper belongs, is an abstract concept. Though there is an available taxonomy of various computer science related topics,[4] annotating a paper with an appropriate set of keywords from such a classification framework is left to the discretion of the author. Therefore, this is not a reliable source from which one can extract the research topic to which a paper belongs. The notion of a research idea remains latent during the interpretation of the content of a paper. One acceptable approach would be to identify a research idea or a topic by grouping a collection of research papers using an unsupervised topic modeling approach like latent Dirichlet allocation [27] and its variants [28]. Recently, several studies [4], [29]–[31] have used LDA based approaches to model research topics from scholarly data. For instance, [31] provides a report on how topics (extracted using LDA) are distributed across authors, publication venues, and citations. Other studies on semantic analysis [32], and collaborative filtering [33] offer insights into the latest results in this area.

Collaboration and interaction among researchers within and across disciplines are the cornerstones of successful research. There exists a significant body of work that has used ideas from network science to characterize interaction among researchers. The seminal work by Newman *et al.* [34], [35] observed the small world phenomenon in collaboration by analyzing the publication data from biomedical science and physics. An early work by Newman [35] has also investigated collaboration among authors and computed various network metrics such as the average path length, degree centrality, clustering coefficient for the author and paper-based networks. The notion of interdisciplinary research has been characterized through co-author networks. Andrade *et al.* [36] discuss various dimensions of collaborations among researchers such as inter and intradisciplinary interactions. Researchers have analyzed inter-disciplinary collaborations of authors [37] in CNRS laboratories. Researchers have also reported empirical evidence of collaboration between organizations [38] from a Belgian manufacturing dataset. Recently, researchers have found that constructing a bibliographic coupling network [39], [40] among published papers can provide interesting insights of the interdisciplinary nature of scientific work. While the work mentioned above analyzes a network, studies like Vivo [41] implements a social networking framework for interdisciplinary collaboration. The work by Ding *et al.* [42] used LDA to extract topics from the publication corpora and analysed the impact of a topic on the collaboration among authors.

Researchers have also developed recommendation systems to suggest collaboration using cross-domain topic modeling [43], [44]. Using citation data, researchers have proposed a future interdisciplinary collaboration model [44]. For a specific domain within computer science discipline, researchers have studied the network characteristics in software engineering research [45] and analyzed various factors influencing research contributions and research collaborations [46].

Researchers have studied how interdisciplinary research is related to scientific impacts based on citation data. Lariviere and Gingras explore relationships between multidisciplinary papers [47]. In fact, they found that highly intradisciplinary and highly interdisciplinary papers attract low citations. Furthermore, the researchers observed a Mathew effect [21] in citation attraction for a paper, if the paper cites previously published papers from highly cited disciplines. A similar work [48] used Scopus database to show that interdisciplinarity of paper has a positive effect on scientific impact. A work by Glanzel *et al.* has chosen the bioinformatics discipline to study interdisciplinary impacts [49]. Another work considers 24 significant subjects and observed relative variances in journal impact measure [50]. A recent work by Dong *et al.* [51] found that the number of citations in a paper has drastically increased due to the interdisciplinary nature of modern science.

B. OBSERVATIONS

From the above overview of related work, we observe that there is strong recognition of the influence of interdisciplinary ideas in shaping the direction of research in recent years. While attempting to understand the influence, existing studies have considered broad disciplines such as physics, biology, computer science etc. These studies have primarily focused at the level of papers, paper citations; and networks of authors, papers, and citations. As a complement to these approaches, our unit of analysis is a research topic, which is at a higher level of abstraction than papers and authors. This abstraction helps us in quantifying the interdomain presence of topics on the basis of textual – rather than citation or collaboration based – similarity measurement [52] and offers a broader range of insights on research ecosystems of the computing domains we have studied.

VI. SUMMARY AND CONCLUSIONS

In this paper, we report result from an examination of the factors that relate to the inter-domain presence of research topics. We examine large corpora of research publications from four domains within the computing discipline: artificial intelligence, databases, operating systems, and software engineering. Using natural language processing techniques, we discover a set of research topics from each domain. On the basis of cosine similarity between topic keywords, we construct a topic network across all four domains. A multiple linear regression model using suitable proxies for inter-domain

[4] https://www.acm.org/about/class/2012

presence of topics, and other known factors that can potentially impact the inter-domain presence is developed and studied.

The overall model is statistically significant and can explain more than 85 percent of the variability in the data. The correlation between actual and model predicted values of the variable representing inter-domain presence of topics is more than 0.9, and almost all model variables have statistically significant effects. Surprisingly, we find evidence that *fewer* number of authors publishing on a topic and a *lower* level of their collective eminence relates to *higher* inter-domain presence of that topic; while *more* papers and venues for a topic, each relates to *higher* its inter-domain presence. The domain a topic belongs to, also has a statistically significant and direct relationship with the topic's inter-domain presence. As expected, topics which connect *many* disparate topics, have a *higher* inter-domain presence while those that *largely* belong to close-knit clusters have a *lower* inter-domain presence.

Our results reveal new motifs in the ecosystem of inter-domain topics. Contrary to conventional wisdom, we find evidence that the involvement of many authors or highly prominent ones, *do not* relate to higher inter-domain presence of topics. Higher inter-domain presence of topics appear to be more closely associated with characteristics inherent to the topics themselves. These results can help individual researchers identify research topics they want to explore and pursue; and facilitate research organizations make informed decisions on proposals and in the governance of research groups.

REFERENCES

[1] T. S. Kuhn, *The Structure of Scientific Revolutions*, 3rd ed. Chicago, IL, USA: Univ. Chicago Press, Dec. 1996.

[2] D. D. S. Price, *Little Science, Big Science...and Beyond*. New York, NY, USA: Columbia Univ. Press, Aug. 1986.

[3] I. Berlin, *The Hedgehog and the Fox: An Essay on Tolstoy's View of History*. Chicago, IL, USA: Ivan R. Dee Publisher, 1993.

[4] S. Datta, S. Sarkar, and A. S. M. Sajeev, "How long will this live? Discovering the lifespans of software engineering ideas," *IEEE Trans. Big Data*, vol. 2, no. 2, pp. 124–137, Jun. 2016.

[5] R. R. Brown, A. Deletic, and T. H. F. Wong, "Interdisciplinarity: How to catalyse collaboration," *Nature News*, vol. 525, no. 7569, Sep. 2015, Art. no. 315. [Online]. Available: http://www.nature.com/news/interdisciplinarity-how-to-catalyse-collaboration-1.18343

[6] L. E. Travis, "In defense of artificial intelligence research," *Commun. ACM*, vol. 5, no. 1, pp. 6–7, Jan. 1962. [Online]. Available: http://doi.acm.org.library.sutd.edu.sg:2048/10.1145/366243.366283

[7] B. Hodjat, "The AI resurgence: Why now?" 2015. [Online]. Available: http://www.wired.com/insights/2015/03/ai-resurgence-now/ Last accessed on: May 21, 2016.

[8] D. Abadi, R. Agrawal, A. Ailamaki, M. Balazinska, P. A. Bernstein, M. J. Carey, S. Chaudhuri, J. Dean, A. Doan, M. J. Franklin, J. Gehrke, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, H. V. Jagadish, D. Kossmann, S. Madden, S. Mehrotra, T. Milo, J. F. Naughton, R. Ramakrishnan, V. Markl, C. Olston, B. C. Ooi, C. Ré, D. Suciu, M. Stonebraker, T. Walter, and J. Widom, "The Beckman report on database research," *SIGMOD Rec.*, vol. 43, no. 3, pp. 61–70, Dec. 2014. [Online]. Available: http://doi.acm.org.library.sutd.edu.sg:2048/10.1145/2694428.2694441

[9] P. J. Denning, "Fifty years of operating systems," *Commun. ACM*, vol. 59, no. 3, pp. 30–32, Feb. 2016. [Online]. Available: http://doi.acm.org.library.sutd.edu.sg:2048/10.1145/2880150

[10] M. Shaw, "Continuing prospects for an engineering discipline of software," *IEEE Softw.*, vol. 26, no. 6, pp. 64–67, Nov./Dec. 2009.

[11] M. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," *Phys. Rev. E*, vol. 64, no. 1, 2001, Art. no. 016131. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.64.016131

[12] M. F. Porter, "An algorithm for suffix stripping," in *Readings in Information Retrieval*, K. S. Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann, 1997, pp. 313–316. [Online]. Available: http://dl.acm.org/citation.cfm?id=275537.275705

[13] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *Ann. Appl. Statist.*, vol. 1, no. 1, pp. 17–35, Jun. 2007.

[14] W. Zhao, W. Zou, and J. J Chen, "Topic modeling for cluster analysis of large biological and medical datasets," *BMC Bioinf.*, vol. 15, no. Suppl 11, Oct. 2014, Art. no. S11.

[15] D. S. Freedman, L. K. Khan, M. K. Serdula, W. H. Dietz, S. R. Srinivasan, and G. S. Berenson, "The relation of childhood BMI to adult adiposity: The Bogalusa heart study," *Pediatrics*, vol. 115, no. 1, pp. 22–27, 2005.

[16] J. Riegelsberger, M. A. Sasse, and J. D. McCarthy, "Shiny happy people building trust?: Photos on e-commerce websites and consumer trust," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2003, pp. 121–128.

[17] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, pp. 167–256, Mar. 2003. [Online]. Available: http://arxiv.org/abs/cond-mat/0303516

[18] S. Datta, P. Basuchowdhuri, S. Acharya, and S. Majumder, "The habits of highly effective researchers: An empirical study," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 3–17, Mar. 2017.

[19] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Nat. Academy Sci. United States America*, vol. 102, no. 46, pp. 16 569–16 572, Nov. 2005.

[20] B. Tabachnick and L. Fidell, *Using Multivariate Statistics*. Boston, MA, USA: Pearson Education, 2007.

[21] R. Albert and A. Barabasi, "Statistical mechanics of complex networks," *cond-mat/0106096*, Jun. 2001, *Rev. Modern Physics*, vol. 74, pp. 47–97, 2002. [Online]. Available: http://arxiv.org/abs/cond-mat/0106096

[22] D. Barron, "The analysis of count data: Overdispersion and autocorrelation," *Sociol. Methodology*, vol. 22, pp. 179–220, 1992.

[23] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton Univ. Press, Nov. 2010.

[24] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Boston, MA, USA: Wadsworth, Jan. 2001.

[25] J. Grudin, "Technology, conferences, and community," *Commun. ACM*, vol. 54, no. 2, Feb. 2011, Art. no. 41. [Online]. Available: http://cacm.acm.org/magazines/2011/2/104400-technology-conferences-and-community/fulltext

[26] R. Van Noorden, "Metrics: A profusion of measures," *Nature News*, vol. 465, no. 7300, pp. 864–866, Jun. 2010. [Online]. Available: http://www.nature.com/news/2010/100616/full/465864a.html

[27] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[28] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[29] Y. Jo, J. E. Hopcroft, and C. Lagoze, "The web of topics: Discovering the topology of topic evolution in a corpus," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 257–266.

[30] B. Hu and C. Zhang, "A lead-lag analysis of the topic evolution patterns for preprints and publications," *J. Assoc. Inf. Sci. Technol.*, vol. 66, pp. 2643–2656, 2015.

[31] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 1055–1060.

[32] Y. Zhang, D. Yi, B. Wei, and Y. Zhuang, "A GPU-accelerated non-negative sparse latent semantic analysis algorithm for social tagging data," *Inf. Sci.*, vol. 281, pp. 687–702, Oct. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2014.04.047

[33] H. Li, K. Li, A. Jiyao, and K. Li, "MSGD: A novel matrix factorization approach for large-scale collaborative filtering recommender systems on GPUs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 7, pp. 1530–1544, Jul. 2018.

[34] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Academy Sci. United States America*, vol. 98, pp. 404–409, 2000.

[35] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proc. Nat. Academy Sci. United States America*, vol. 101, pp. 404–409, 2004.

[36] H. B. Andrade, E. de Los Reyes Lopez, and T. B. Martin, "Dimensions of scientific collaboration and its contribution to the academic research groups? Scientific quality," *Res. Eval.*, vol. 18, pp. 301–311, 2009.

IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

[37] A. Sigogneau, O. Malagutti, M. Crance, and S. Bauin, "Cross-disciplinary research: Co-evaluation and co-publication practices of the CNRS laboratories," *Res. Eval.*, vol. 14, pp. 165–176, 2005.

[38] R. Veugelers and B. Cassiman, "R&D cooperation between firms and universities. Some empirical evidence from Belgian manufacturing," *Int. J. Ind. Org.*, vol. 23, pp. 355–379, 2005.

[39] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, pp. 1386–1409, 2010.

[40] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, pp. 1313–1326, 2012.

[41] V. Gewin, "Networking in vivo: An interdisciplinary networking site for scientists," *Nature*, vol. 462, 2009, Art. no. 123.

[42] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *J. Informetrics*, vol. 5, no. 1, pp. 187–203, 2011.

[43] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1285–1293.

[44] Y. Guo and X. Chen, "Cross-domain scientific collaborations prediction using citation," in *Proc. IEEE/ACM Int. Conf. Advances Social Netw. Anal. Mining*, 2013, pp. 765–770.

[45] S. Datta, N. Kumar, and S. Sarkar, "The social network of software engineering research," in *Proc. 5th India Softw. Eng. Conf.*, 2012, pp. 61–70.

[46] S. Datta, A. S. M. Sajeev, S. Sarkar, and N. Kumar, "Factors influencing research contributions and researcher interactions in software engineering: An empirical study," in *Proc. 20th Asia-Pacific Softw. Eng. Conf.*, 2013, pp. 34–41.

[47] V. Lariviere and Y. Gingras, "On the relationship between interdisciplinarity and scientific impact," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, pp. 126–131, 2010.

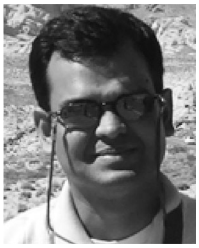
[48] F. Davletov, A. S. Aydin, and A. Cakmak, "High impact academic paper prediction using temporal and topological features," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 491–498.

[49] W. Glanzel, F. Janssens, and B. Thijs, "A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics," *Scientometrics*, vol. 79, no. 1, pp. 109–129, 2009.

[50] K. W. Boyack and R. Klavans, "Predicting the importance of current papers," in *Proc. 10th Int. Conf. Int. Soc. Scientometrics Informetrics*, 2005, pp. 335–342.

[51] Y. Dong, H. Ma, Z. Shen, and K. Wang, "A century of science: Globalization of scientific collaborations, citations, and innovations," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1437–1446.

[52] D. R. Amancio, M. G. V. Nunes, O. N. Oliveira Jr, *et al.*, "Using complex networks concepts to assess approaches for citations in scientific papers," *Scientometrics*, vol. 91, no. 3, pp. 817–842, 2012.



SUBHAJIT DATTA has two decades of experience in software design, development, research, and teaching at various organizations in the United States of America, India, and Singapore. He is currently an assistant professor (education) in the School of Information Systems, Singapore Management University. More details about his background and interests are available at www.dattas.net.

RUMANA LAKDAWALA received the bachelor's degree in computer science from BITS Pilani, Goa. She is currently working toward the master's degree in artificial intelligence at KU Leuven, Belgium. Her interests include big data analysis, machine learning, and text based information retrieval.

SANTONU SARKAR is a professor of computer science and information systems, BITS Pilani, K.K. Birla Goa Campus. His current research interests include building software engineering techniques to ensure dependability, performance, and ease-of-use of cloud, cyber-physical, and HPC applications. More details are available at http://www.bits-pilani.ac.in/goa/ComputerScienceInformationsSystems/software.