Accessibility metric for characterizing the relevance of conference papers

Paramita Das
Dept. of CST
IIEST, Shibpur
Howrah, India
dasparamita1708@gmail.com

Agniv Adhikari
Computer Division
CSIR-CGCRI
Kolkata, India
agniv.adhikari@gmail.com

Abhik Mukherjee
Dept. of CST
IIEST, Shibpur
Howrah, India
abhikm.kol@gmail.com

***Abstract*—Academic conference is a medium for rapid dissemination of knowledge. In an expanding horizon, often it becomes difficult to achieve proper impact due to topic mismatch. The problem affects either way, both conference organizers as well as the prospective participants. As a result, some of the works remain disconnected and disoriented from the majority volume of work presented in the conference. A solution to such situation is sought here by establishing better cohesion. The papers are modelled as nodes of a graph and these are connected through edges if they share a common keyword, specified during submission. An accessibility metric over this network is proposed in this work as being capable of judging the relevance of a paper. Two case studies are presented for proof of concept.**

***Index Terms*—network theory, text mining, graph metric, accessibility, keyword based graph**

I. Introduction

In coordination with the progress of research thoughts, conferences are arranged, journals are published to accumulate the research articles in a structured way. Although both being the publication medium of research documents, the characteristics of conference are a bit different from journal- conferences are more dynamic and heterogeneous in nature. A typical conference series is a variant combination of research domains which may vary periodically. Depending on the venue of the conference being organized, there may be an issue of special call for papers. Moreover the prospective author list is also influenced by the choice of venue. So the profile of randomness is a matter of observation in case of conferences. Regarding these facts, authors and organizers need to be specific in choice of conferences and papers respectively to enhance the flow of knowledge dissemination.

This notion is depicted by the structural information of the network, constructed by the research papers of a specific conference. Each paper consists of standard keywords. Nodes are the papers presented in the conference and edges represent the connectivity of nodes if the papers share common keywords. The interaction between the structure and dynamics is crucial [1], [2]. We propose a metric, accessibility that encodes the relationship. Here the concept of entropy is utilized to quantify the behavioural randomness of a node, i.e. paper of the network.

Accessibility is a measure used in transit system to describe how easy it is for a system to facilitate the opportunity of access. There is a number of possible ways to define accessibility in traffic network [3]. Among them entropy based approach [4], [5] is one which has been employed here. The metric correlates the structural diversity with the variation of entropy. In our context higher the value of entropy– greater the randomness, i.e, accessibility of the paper and vice versa. Random walk [6], [7], [8], [9] is an obvious technique to model the probability distribution for measuring variance signature of entropy, specially in the context of networks [10].

Due to inconsistency in topic similarity some works remain disoriented from the majority volume of the work presented in the conference. Our aim is to find those papers by introducing the suitability measure for a conference series. The remainder of the paper is organized as follows: Section II describes the proposed metric to quantify accessibility measure with a suitable example. The implementation details like data collection and data representation are provided in Section III. The results in the form of two case studies are described in Section IV. The necessary conclusions, interpretations, and further directions are discussed in Section V.

II. Methodology

A. Network of research papers

Papers presented in the conference form the nodes of a graph. It is assumed that for each paper, a number of keywords have been designated from a standard vocabulary of IEEE. Nodes are connected through edges if they share the same keyword. This approach has been successfully studied in [11]. The edge weight is given as an integer that represents the count of the number of common keywords between two papers. If random walk is initiated from a node, only few nodes can be reached in one or more hop(s). Likewise, random walks from few nodes can terminate in a given node in one or more hop(s). This can give a measure of the knowledge dissemination factor pertaining to a paper, resembling how it interacts or reaches out to other papers in the conference in terms of knowledge. The metric of outward accessibility has been borrowed from the field of traffic networks to define a suitable metric for the papers. Average accessibility and departure of individual papers from this average value can give the idea about the homogeneity among all papers presented in the conference.

978-1-7281-0297-9/19/$31.00 ©2019 IEEE

Generally conferences are held periodically. The relevance of keywords has some sensitive dependence on the venue. Nevertheless, the relevance of a prospective submission to an upcoming conference can be judged by computing its accessibility in the past conferences of this series- simply by inserting it as a node and drawing edges based on its chosen keywords. The computed accessibility can then be compared with the average accessibility computed for the past conference to get some relative idea about how it would have fared in that conference. Likewise, the conference organizers may also compute an accessibility metric by forming a graph based on papers selected for their upcoming conference. This can provide them with an estimate of the likely interactions in the upcoming conference.

Let a graph $G = (V, E)$ be given where V is the set of collected conference papers, whose elements act as the nodes of the graph. Each paper $v_i \in V$ contains more than one keywords, i.e. k_{ij} and $j \in \{1, 2, .., n\}$. If at least one keyword is common to any of the two papers of the graph, then there exists an edge between the two papers, i.e. nodes. For commonality of two different keywords in between nodes v_i and v_j, two parallel edges exist between node v_i and v_j. The number of parallel edges between any two nodes is provided as the weight of the link between the two nodes.

B. Accessibility metric

Network is the essential model that encompass the complex relationships of most of the diverse systems and random walk is such a stochastic process that captures this diversity. The random walk can be interpreted as– out of total probable paths– how the destination differs that depends on the structure of the system. We try to capture this concept in our context of research media.

Now the self-avoiding random walk is defined as– starting from a node v_i the walk chooses nodes among the neighbours of the current node uniformly at random and touches the nodes as many as possible at a distance of h without revisiting the previous nodes. So a self-avoiding random walk of length h is the sequence of $h + 1$ nodes and h edges but the none of the nodes or edges are repeated.

Let the total number of possible self-avoiding walks starting from node i and preceding h steps be N where $h \in \{1, 2, ..., H\}$. Now, the link probability that a self-avoiding walk arrives at node j starting from node i after h steps is denoted as $P^h(j, i)$ and M be the total number of walks possible from node i to node j preceding h steps. So the probability at hop h can be expressed as-

$$P^h(j, i) = \frac{M}{N} \tag{1}$$

The walk may stop when any of the following three criteria is satisfied:

- The walk has already covered maximum predefined steps or hops H.
- The walk reaches to a pendant node of the graph.
- All neighbouring nodes have been traversed by the walk.

Now, the entropy of the node i is measured by the Shannon's entropy as-

$$E^h(i) = -\sum_{j=1}^{|V|} \begin{cases} 0 & if P^h(j,i) = 0, \\ P^h(j,i) log(P^h(j,i)) & if P^h(j,i) \neq 0, \end{cases} \tag{2}$$

So, the outward accessibility of node i at hop h becomes-

$$A^h(i) = \frac{\exp(E^h(i))}{|V| - 1} \tag{3}$$

C. Illustration

The accessibility of node A in the sample graph is computed below:

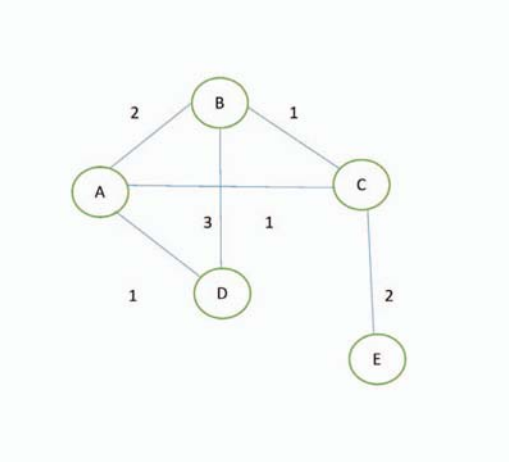

Fig. 1. Example graph for calculation of accessibility

All the self avoiding random walks starting from node A to remaining nodes are tabulated below:

TABLE I
HOP WISE SELF AVOIDING RANDOM WALKS

hop	paths
1	$[A \to B] = 2$ $[A \to C] = 1$ $[A \to D] = 1$ $[A \to E] = 0$
2	$[A \to C \to B] = 1$ $[A \to D \to B] = 3$ $[A \to B \to C] = 2$ $[A \to B \to D] = 6$ $[A \to C \to E] = 2$ and all the paths of hop 1
3	$[A \to B \to C \to E] = 4$ $[A \to C \to B \to D] = 3$ $[A \to D \to B \to C] = 3$ and all the paths of hop 1 and 2

Now the probability of all the paths from node A to other nodes are calculated following the Equation 1:

TABLE II
HOP WISE PROBABILITY OF THE LINKS

hop	$P(B,A)$	$P(C,A)$	$P(D,A)$	$P(E,A)$
1	0.5	0.25	0.25	0
2	0.33	0.17	0.39	0.11
3	0.21	0.21	0.35	0.21

Finally the entropy and accessibility are computed following Equation 2 and 3 of each hop for the node A:

TABLE III
HOP WISE ENTROPY AND ACCESSIBILITY

Hop	Entropy	Accessibility
1	1.03	0.70
2	1.26	0.89
3	1.35	0.97

III. IMPLEMENTATION

A. Data collection

We have chosen IEEE website to collect the papers for its vast and variant assembly of research articles. Moreover the papers presented in a conference hold DOI (Digital Object Identifier) which are sequential and unique. So it becomes easy to program a web crawler to download the papers following the sequential DOIs.

To develop the crawler, a property file has been generated which contains the two following things–

- First paper DOI of a particular conference for a year.
- Paper count of the conference for that year.

Now the file is passed to the crawler. It starts with the base URL and the first DOI and then increments the DOI in a loop to download the specified number of papers. Along with the papers we aim to get the IEEE keywords also. For this purpose the string *keywords* is supplied to the URL and DOI to obtain the IEEE keywords associated with each paper. After completion of the download process, we extract the keywords using regular expression.

B. Data Representation

Each of the papers of a particular year are encoded into nodes of a graph. At the very beginning, all the nodes were isolated and gradually they form a network by setting up links using the keyword information. Once the network has been built, the accessibility metric is computed for 3 hops. So the following data files are required.

1) A dictionary of unique paper id, i.e. DOI and its associated keywords list.
2) A dictionary of IEEE keywords and the list of papers in which the particular keyword has occurred. Here each keyword is treated as the key of the dictionary.
3) Adjacency list of each paper for a particular year. As output files, for each paper computed accessibility for each of the hops are collected into a year wise matrix.

IV. RESULTS

Results obtained are based on two regular annually held IEEE conferences. The choice of conferences is such that one is a conference that has focus on specific area and the other is a symposium covering a vast spread of domains. Keywords are chosen to be same as those declared by the authors as well as by IEEE. Outward accessibility metric is computed over different years to see the trend emerging out of the analysis.

A. Domain specific onference

As a concrete example, IEEE I2MTC has been chosen. For this flagship instrumentation and measurement technology conference series, each paper presented in 2017 is placed as pseudo-node in 2016 graph separately and accessibility upto 3 hops is computed and compared with their actual value for 2017. It is reasonable to assume that the reachability of prospective audience of a conference paper would be restricted to 3-hop neighbours. Altogether 340 papers were presented in the 2017 edition of the conference, IEEE suggested keywords are mainly considered and sometimes in case of inadequacy, author defined keywords are supplemented. Altogether 11 papers were in any case omitted from the analysis due to inadequate number of available keywords.

It has been found that the accessibility metric shows consistent result– the accessibility of papers in real network of 2017 remain more or less in the same range for 2016 graph. The obtained accessibility values are normalized with respect to the hopwise maximum and minimum values of accessibility in respective domains. The average accessibility has been computed for all three cases, i.e. actuals of 2016 and 2017 as well as the projection, and is tabulated in Table IV for ready reference.

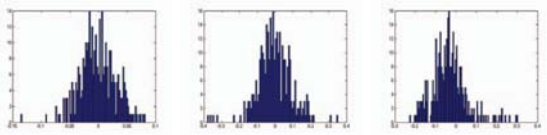

Fig. 2. Error histogram of I2MTC for first 3 hops (L-Hop1, C-Hop2, R-Hop3)

The mode of the error histogram of the year 2017 and projected in the year 2016 for all 3 hops stays close to zero as evident from Fig. 2. Since the $3-hop$ average accessibility of 2016 is higher than that of 2017 (vide Table IV), the error histogram mode appears slightly to the left $(-ve)$ of zero. Evidently for larger hops, the number of papers which show high difference between the computed accessibility for 2017

TABLE IV
HOP WISE AVERAGE ACCESSIBILITY

I2MTC	Hop-1	Hop-2	Hop-3
Actuals of 2017	0.1111	0.5304	0.6810
Actuals of 2016	0.1145	0.5234	0.7101
Projected on 2016	0.1131	0.5359	0.7165

and the projected accessibility for 2016 becomes less, but the decrease is not too drastic. As the number of hops increase, there would be a tendency of saturation in count of possible random walks with respect to the reachability from a node.

The top five cases where the accessibility difference was drastic were analyzed individually. Out of those cases where the $3 - hop$ accessibility of 2017 was on higher side, there were 3 cases where the work was on fault tolerant techniques in manufacturing, a specialized topic that did not find matches in 2016. One case belonged to speaker recognition system which seemed slightly misplaced and another one was related to bio-medical ultrasonics where the authors used abbreviated keywords as well as the topic is specialized and that may have reduced the accessibility when projected on to 2016.

The top five cases where the 2016 projected accessibility stayed on higher side included 2 very specialized papers on bio-medical sensors, one on geo-technical systems, one on rehabilitation techniques and another on faults in photo-voltaic panels. These papers did not fit well in the tracks of 2017 but used keywords that found better company in 2016 edition.

From this fault analysis, it may be concluded that the accessibility metric has grossly performed correctly, barring few cases of topic mismatch and few cases of wrong keyword usage. This also leads to the scope of revisiting keyword suggestions for better interaction of a paper with its neighbourhood. Low accessibility can often be attributed to papers on emerging areas or special tracks. In such cases, the metric could be misleading. Organizers need to judge case-by-case and decide on inclusion of low accessibility papers.

B. Conference with vast domain coverage

As an example of such conference, the IEEE Tensymp has been chosen. For this conference, the metadata of papers presented in 2016 and 2017 have been collected. The analysis conducted follows a similar line as above. It has been found that the individual accessibility values of the papers remain smaller in general than those in case of the other case study. This seems consistent since this conference coverage is more so that a number of papers related to some topic remains weakly connected with those from another topic. The average values are tabulated in Table V. The mode of the error histogram of the year 2017 and projected in the year 2016 for all 3 hops stays close to zero as evident from Fig. 3.

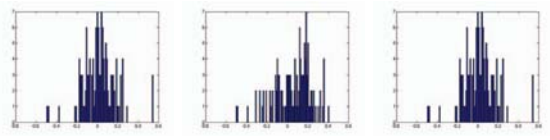

Fig. 3. Error histogram of TenSymp (L-Hop1, C-Hop2, R-Hop3)

The broad coverage ensured that the number of inaccessible papers remained small. Nevertheless, few papers could not match in terms of the accessibility metric. Papers of 2017 that found better accessibility when placed in 2016 included topics like mobility for campus network, automotive engineering, cyber attacks on power system, relay network location and cloud resource management. Actual 2017 papers that had high accessibility but did not fit well in 2016 included topics like electrocardiography, smart phone signal attenuation, emergency services for smart city, middleware for smart city.

TABLE V
HOP WISE AVERAGE ACCESSIBILITY

TENSYMP	Hop-1	Hop-2	Hop-3
Actuals of 2017	0.1098	0.4150	0.6247
Actuals of 2016	0.2085	0.5330	0.6309
Projected on 2016	0.0902	0.4736	0.6430

V. CONCLUSION

In this work, accessibility metric has been defined for conference papers. This metric indicates how well the paper intermingled with other papers presented in the same conference. Good match in accessibility has been found between papers actually presented and projected accessibility for the preceding year. It has been observed that with increasing hops, the accessibility metric becomes insensitive and loses the purpose. Hence the present study has not been extended beyond 3 hops. Detailed analysis of discrepancies in accessibility metric has also been conducted. The reasons for such discrepancy could be explained properly from the domainm knowledge about the conference. Besides, from the analysis we can have a clear picture of the new emerging topics for future conferences.

One limitation is that the author supplied keywords are used for the purpose. In order to extend the work to news media streams, some keyword extraction algorithm needs to be used in absence of such author defined keywords.

REFERENCES

[1] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek, Evolution of the social network of scientific collaborations, vol. 311., Physica A: Statistical mechanics and its applications, 2002, pp. 590–614.

[2] Dan Braha, Complex Design Networks: Structure and Dynamics, arXiv preprint arXiv:1801.02272, 2018.

[3] SR Jones, Accessibility measures: a literature review, Transport and Road Research Laboratory (TRRL), 1981.

[4] Sven Erlander, Accessibility, entropy and the distribution and assignment of traffic, 3rd ed., vol. 11. Transportation Research, 1977, pp. 149–153.

[5] B.A.N. Travenolo, L.da F. Costa, Accessibility in complex networks, 1st ed., vol. 373. Physics Letters A, 2008, pp. 89–95.

[6] L. Lovász, Random walks on graphs: a survey, 1st ed., vol. 2. Combinatorics, Paul erdos is eighty, 1993, pp. 1–46.

[7] Naoki Masuda, Mason A.Porter, Renaud Lambiotte, Random walks and diffusion on networks, vol. 716-717. Physics Reports, 2017, pp. 1–58.

[8] Chen Avin and Bhaskar Krishnamachari, The power of choice in random walks: an empirical study, Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems, 2006, pp. 219–228.

[9] Kartik Anand and Ginestra Bianconi, Entropy measures for networks: toward an information theory of complex topologies, 4th ed., vol. 80, Physical Review E, 2009, pp. 45–102.

[10] Roy Timo and Kim Blackmore and Leif Hanlen, On entropy measures for dynamic network topologies: limits to MANET, Proceedings of Communications Theory Workshop, 2005, pp. 95–101.

[11] Agniv Adhikari and Paramita Das and Abhik Mukherjee, Generating a representative keyword subset pertaining to an academic conference series, accepted in Scientometrics, 2019.