Journal of Computer Science

Original Research Paper

An On-line Analytical Processing (OLAP) Aggregation Function for Rising Aspects in Collaboration and Social Networks

[1]Waqas Nawaz, [2]Kifayat Ullah Khan and [1]Muhammad Shoaib Siddiqui

[1]Department of Computer and Information Systems, Islamic University of Madinah, Al-Madinah, Saudi Arabia
[2]Department of Computer Science (FDS Lab), National University of Computer and Emerging Sciences, Islamabad, Pakistan

Article history
Received: 11-03-2020
Revised: 20-05-2020
Accepted: 09-06-2020

Corresponding Author:
Waqas Nawaz
Deptment of Computer and Information Systems, Islamic University of Madinah, Al-Madinah, Saudi Arabia
Email: wnawaz@iu.edu.sa

Abstract: The overwhelming usage of social and collaboration networks provides the opportunity to analyze evolution of relationships among individuals, like celebrities or co-authors. Discovering such phenomenon in large complex networks is non-trivial due to their large sizes. In this situation, the aggregation functions used in OLAP, are useful to analyze the summarized data. OLAP has successfully proven its worth on multidimensional or complex networks. However, existing aggregations in the current OLAP systems do not produce versatile results in case of social and collaboration networks. This happens because said type of networks have structural connectivity/links among nodes, which cannot be considered by OLAP during its execution. In this situation, a useful discovery in terms of identifying pairs of nodes whose relationships is emerging in recent time, is missed. Such discovery of pairs of nodes is important for various applications such as targeted marketing, future joint partnerships and predicting future correspondence to name a few. In this study, we call such pairs as Rising_Pairs and propose an aggregation function for performing OLAP on network data whose historical information is maintained over a period of times. Using structural information, Rising_Pairs, our proposed aggregation function, discovers the strongly coupled pairs in a network data by emphasizing their recent interactions and attribute similarities. In this way, useful information related to strongly coupled pairs in a network is identified. To verify the effectiveness of our proposal, we implemented it on various types of real-world networks like Facebook, Digital Bibliography and Library Project (DBLP) and Global Positioning System (GPS) trajectory datasets and observed interesting patterns.

Keywords: Social Networks, OLAP, Aggregation Functions, Data Analysis, Bibliographic Data, Rising Phenomenon, Temporal Trends, Summarization

Introduction

Online social networks are extremely popular these days. People are making excessive use of them for wide variety of purposes such as social interaction (Huberman *et al*., 2008), content sharing (Lange, 2007), community detection (Li and Shen, 2011), viral marketing (Subramani and Rajagopalan, 2003), news sharing (Li Zhang *et al.*, 2004), job recruitment (Calvo-Armengol and Jackson, 2004) and many others. All such tasks are easily performed, hence, the number of users is increasing rapidly. Similarly, many social media users are interested to analyze interactions (in the form of pairs) between the celebrities, they follow. The celebrities can be of type couples of movie actors and actresses or a pair of researchers collaborating together. It is of keen interest of such users to see how the individual relationships and the personal interests of some of the celebrities, are changing with the passage of time. They want to see how their favorites got close to each other and adopted similar habits although their initial relationships and similarities were too little. In a nutshell, the recent and growing affiliations of the celebrities, are important to be analyzed compared to their interactions in the past.

To identify aforementioned types of scenario in the domain of bibliographic data, we find that various tools and systems have been developed in literature like Bernabei *et al*. (2015); Burch *et al*. (2015); Pflugrad (2017); Mezzanzanica *et al*. (2018); Cesarini *et al*. (2018); Mercorio *et al*. (2019a). Modelling data of DBLP as a social network and then performing various network analysis tasks on this dataset is an interesting area of research as highlighted by Biryukov and Dong (2010). Such analysis on DBLP provides answers to questions such as analyze the research community in the computer science field Babskova *et al.* (2013) Kumar *et al.* (2017) Cabot *et al*. (2018) Mercorio *et al*. (2019b) Abazi-Bexheti *et al*. (2019), identifying the field experts Yang *et al*. (2013) Moreira *et al*. (2015) Pflugrad (2017), publication and venue quality analysis Ueda *et al*. (2017) Fathalla *et al*. (2018) Herrmannova (2018) Keselman (2019) to name a few. An interesting scenario is when the researchers had co-authored more papers at the start of their collaboration period and have similar research areas. However, there is no joint contribution from them in the recent past and their research interests have also changed. This pair can be regarded as strongly bonded in terms of having large number of joint publications but if their recent interactions are low, then they cannot be titled as tightly connected pairs.

Analyzing interactions among pairs of nodes (referred as celebrities in previous paragraph), over a period of time, in large social networks is computationally expensive (Tang *et al*., 2009). However, quantitative analysis using various aggregation functions provide useful statistics in an efficient manner. In this regard, OLAP is a useful database tool. Using OLAP, we can perform various analytics like roll-up, drill down, slice and dice. OLAP also makes highly use of various aggregation functions to analyze the summarized data at various levels of granularity.

We observe that many researchers are using OLAP techniques in social and collaboration network datasets (Zhao *et al*., 2011; Chen *et al*., 2008; Qu *et al*., 2011; Queiroz-Sousa and Salgado, 2019; Bleco and Kotidis, 2019; Ghrab *et al*., 2020). On the other hand, applying aggregation functions during OLAP ignore the structural connectivity among nodes. The exiting aggregation functions produce only the simple summarized results like total number of nodes, maximum edge weight, path length among others. However, analyzing phenomenon like increasing and decreasing trends over the time, cannot be easily captured by them.

In this study, we propose an aggregation function, Rising_Pairs, for performing OLAP on network datasets that estimates the strength of relationship between the pair of nodes over a given period of time. The proposed function discovers the pairs of nodes in a network whose initial coupling was weak, but they are emerging tightly coupled pairs. To identify such pairs, we first build timed stamped construction of social and collaboration networks to maintain the historical interactions in edge attributes between the nodes and their interests/habits as attributes. Rising_Pairs then aggregates the interactions and attributes similarities to find required pairs of nodes. The underlying phenomenon of our proposed function is based on a statistical measure Exponential Moving Average (EMA) (Lawrance and Lewis, 1977) and is published as our conference paper (Khan *et al*., 2012).

Related Works

OLAP is a useful tool to analyze aggregated data at various levels of granularity. From its success stories on relational data, people have used for variety of data like unstructured (Baars and Kemper, 2008), sequence (Lo *et al*., 2008), streaming (Han *et al*., 2005) and many others. OLAP for social networks is another much focused area in the recent past. Zhao *et al*. (2011; Chen *et al*., 2008; Qu *et al*., 2011) the data warehousing and OLAP frameworks are presented for social networks and is comprehensively explained how to take maximum benefit from such decision support tools in this scenario. All of these research studies describe measures as the aggregated graphs, but their focus is not towards utilizing the aggregation functions on the underlying data. Similarly, OLAP is used as a concept in some research studies on graphs to analyze the data at various levels of details. For example, Tian *et al*. (2008) provides graph summarization using SNAP and k-SNAP operations but resembles much with applying clustering on graphs. Similarly the objectives of graph summarization and compression techniques in (LeFevre and Terzi, 2010; Liu *et al*., 2008; Navlakha *et al*., 2008) are significantly different to that of ours.

The motivation of the proposed function, Rising_Pairs, also resembles to finding the closely related vertices in graphs. In this respect, hierarchical clustering (Kaufman and Rousseeuw, 2009) is most similar. However, Rising_Pairs concentrates on the historic relationship pattern of directly connected vertices, analyzes the recent similarity of attributes values. The graph theoretical concept of betweenness centrality is also related but it differs as it helps finding influential nodes in the graph and more inclined towards shortest path problem (Freeman, 1977) and so is the case of measures like vertex between ness and edge between ness (Girvan and Newman, 2002).

The OLAP aggregation functions have shown promising results on relational data. However there are few studies like (Chui *et al*., 2010; Ravat *et al*., 2008; 2007) in which the authors have proposed new aggregation functions to perform OLAP on sequence and text data as the existing functions do not satisfy the domain specific requirements.

The concept for historical organization of social networks is related to the model, Time aggregated graph, presented in (George and Shekhar, 2008). The elements of the graph are attached with time varying attributes to capture the values over the passage of time which are quite similar to the edge attributes of the graph organization presented in this study. However, time aggregated graph is designed specifically keeping in mind the requirements of spatiotemporal networks. On the other hand, the motivation for timed historical network is from OLAP perspective whose main task is to support historical analysis. Furthermore, we propose the aggregation of social network's data at various levels of granularity. The focus of this research is the aggregation function which requires the underlying network data in historical manner. We are in the process of delivering more intensive research work on timed historical network in near future. The main idea of contact network presented in (Shirani-Mehr *et al*., 2012) also show some similarities to the timed historical network but differs in such a way that it utilized Time Expanded Network (TEN) as its basis for modeling which vary from timed historical network significantly.

The Graph OLAP model presented in (Chen *et al*., 2008) resembles to the timed historical network as well but differs in such a way that there is no specification to capture the time varying aspects. On the other hand, the timed historical network provides comprehensive details to set the basis for historical analysis with respect to the interaction among the users of the social networks.

Temporal publication trend analysis Song *et al*. (2014) Orr and Ortiz (2013) Swaraj and Manjula (2016) Seo *et al*. (2020) Ryu (2020) is another area of research where authors have used DBLP and similar datasets. In Kim *et al*. (2012), the authors analyze the pattern of publications which eventually help them to identify the scientific output of the research groups. The evolution of communities of researchers with similar topics and interests over a selected time frame is done in Song *et al*. (2014). A very recent study shows that the single author publication are decreasing with the passage of time and discussed various aspects in their study Ryu (2020). The quality and quantity of publications from a selected group over a time span of around eight years is also evaluated for decision making purpose Seo *et al*. (2020).

Historical Organization of Social Networks

The current snapshot or view of the social network explains the present picture such as number of available persons, current value of their attributes and total number of interactions between any two persons. It is unable to show how the network has evolved with the passage of time, what was the attribute value at given time and what was the interaction strength of an arbitrary pair in a given time.

Organizing the historical view of a social network is suitable to perform the trends analysis. The monitoring of individual relationship patterns and behavioral changes can be accomplished only if historical data is available. Therefore, the analysis using the present-day data yields non-context aware results. Moreover, the motivation is to answer any kind of future unknown requirements for analysis.

Time-stamped attributes can be used to maintain all the historical changes of social networks. As the slowly changing dimensions operate in dimensional modeling (Kimball and Ross, 2011), the time-stamped attributes can also be adopted to capture the changes taking place in social networks. This organization supports the historic and trend analysis vital for deriving more knowledge and decision making. Since, it clearly depicts how the two persons have been interacting with each other over a particular time frame.

Figure 1 (a) illustrates the DBLP co-authorship network at certain instance of time. Each vertex denotes an author having an attribute showing his/her research area. And each edge represents the number of co-authored papers. However, it cannot show their recent collaborative work and previous research interests. On the other hand, in Fig. 1 (b), we can easily analyze the historical co-authorship progress and the research interests.

The aggregated data along with the historical information provides a more robust environment to analyze the data at various levels of granularity and from multiple dimensions. There exists an entirely different variety of information at each granular level, which provides more insights of the underlying data. Figure 1 (c) shows the historical aggregation of Fig. 1 (b). There is a trade-off between the aggregation and structural information at the aggregated level. It becomes difficult to maintain the complete structural semantics of the network at higher levels of hierarchy, as there is an essential need to keep the historical information.

Now we present formal definitions for social network, the time stamped attributes and the historically organized social network to clearly communicate our aim in this study.

Definition 1. Social Network (SN)

At given instance of time, say T_i, a social network SN is defined as a network $SN = (V, E, VA, EA)$ where V is the set of vertices ($\forall v \in V$), E is the set of edges ($E \in V_i * V_j$). VA and EA are the attributes attached to vertices and edges respectively.

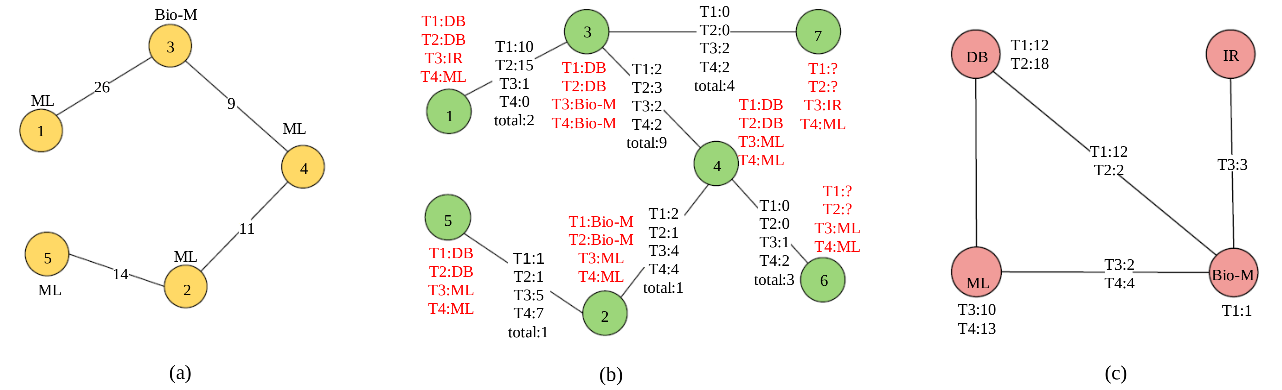

Fig. 1: Sample DBLP co-authorship network at various instances of timestamp, where ML refers to machine learning and bio-m stands for bio medical sciences (a) Sample snapshot at certain; (b) Historical organization at later time; (c) Historical and Aggregated

Definition 2. Vertex Centric Time-Stamped Attributes (VTSA)

Given a set of time intervals $T = T_1, T_2,..., T_n$, *VTSA* is defined as an attribute $a_i \in A$ of vertex V having values $(v_1, v_2,..., v_n)$ for each interval of time.

Definition 3. Edge-Centric Time-Stamped Attribute (ETSA)

Given a set of time intervals $T = \{T_1, T_2,..., T_n\}$, *ETSA* is defined as an attribute $a_i \in A$ of edge E where $(E \in V_i * V_j)$ for each pair of vertices. It stores the count of interactions between a pair of vertices $(u, v) \in V$ at every time interval T_i.

Definition 4. Timed Historical Network (THN)

Given a set of time intervals $T = \{T_1, T_2,..., T_n\}$, *THN* is a collection of social network snapshots $\{SN_1, SN_2,..., SN_n\}$ at each time interval T_i. THN is defined as a network $THN = (V, E, VTSA, ETSA)$ where V is the set of vertices $(\forall v \in V)$, E is the set of edges $(E \in V_i * V_j)$, *VTSA* is the time stamped vertex attribute and *ETSA* is the time-stamped edge attribute.

Algorithm 1 is the pseudo code for constructing THN. The basic idea is to integrate all the snapshots of a network obtained at various time intervals to preserve all the topological changes and changes in attributes of nodes. The algorithm takes the snapshots of the network as various intervals of time and the optional list of dimensions through which aggregations are done. At line 1, the while loop iterates through each snapshot of the network. Lines 2-15 operate on the vertices of $snapshot_i$. The comparisons for aggregation and procedures at Line 3,4 and 13 are only performed when the aggregation is required. Lines 5-7 are the vertex existence comparisons and the corresponding steps are performed and attribute values in *VTSA*. Lines 16-22 operate on the edges for (u, v) in $snapshot_i$. The edge is created along with its *ETSA* on line 18 if it does not exist. Otherwise the interaction count between a pair for snapshot T_i is stored in *ETSA* at line 21.

Proposed Methodology

In this section, we present the proposed aggregation function Rising_Pairs. We begin with its explanation followed by an illustrative example, then discuss the algorithm and finally the performance concerns.

OLAP Aggregation Function: Rising_Pairs

The purpose of the proposed function, Rising_Pairs, is to filter out the strongly coupled pairs of vertices in THN in terms of their higher recent interactions in ETSA and more similarity in the recent values of VTSA. There is more emphasis on the recent behavior rather than on the overall or aggregated one. This is so because there is the possibility that the aggregated value of ETSA for one pair is greater than that of other, but the recent values of former can be much less than those of the later. Similarly, there can be more similarity between the recent VTSA values of one pair than those of the other. A downfall in the recent ETSA values and dissimilarity in the recent VTSA values may mean an ending relationship. Therefore, Rising_Pairs are the pairs of vertices whose recent interaction count is higher and have significant similarities *i* their attribute values.

The motivation for Rising_Pairs is from the statistical function EMA which is a kind of moving average, but more weight is given to the latest data. EMA is a well-known function in businesses, financial circles and stock markets. It gives more weight to the recent trends in the market. We utilize this concept in social and collaboration networks organized as THN and find the rising pairs by considering their recent behavior. The original formula of *EMA* is not suitable to serve the needs of social networks as it incorporates the structural information.

Algorithm 1: Historical data organization of social networks

```
Input: Snapshots of the Social Networks at
       various time intervals, List of Dimensions
Output: Timed Historical Network, THN
while total number of snapshots do
    for each (u, v)∈V do
        if Dimensions of (u, v) are same with respect to
        given Dims then
            create an aggregated vertex av;
            if av does not exist in THN then
                create a vertex and store the current
                Dim values in VTSA;
            end
            else
                add current dimension values in
                corresponding VTSA for time $T_i$;
            end
        end
        else
            create new aggregated vertex and edge for
            every combination of given Dims;
        end
    end
    for each e (u, v) do
        if e does not exist in THN then
            create an edge e and store the current
            interaction count in its ETSA;
        end
        else
            add the current interaction count in
            corresponding ETSA for time $T_i$;
        end

    end
end
```

In order to identify the rising pairs, there is a need to have some measurable criteria to compare and determine the relationship strength of the pairs. We term it as the Rising_Value of each pair. So, the pairs having higher or comparable Rising_Value to that of the pair having the maximum interactions count and most similar characteristics are regarded as the rising/emerging pairs in the THN. The formal definition of Rising_Value is Rising_Value (*ET SA*, $VTSA_u$, $VTSA_v$, *N*) Return [rising value of given pair]. The function receives as input the historical interactions pattern between a pair in ETSA, the historically organized attributes, VTSA, of each user and the total number of snapshots N. It returns the relationship strength of the given pair. The specification for Rising_Value is given in Equation. 1:

$$Rising\_value = \sum_{i=1}^{N}\left(\frac{(ETSA)_{u,v}^{i}}{(N-i)+1} + Sim\left(VTSA_{u}^{i}, VTSA_{v}^{i}\right)\right) \quad (1)$$

where, $ETSA^i$ is the value interaction count or edge weight between the pair is at time T_i, *N* is the total number of snapshots of the *THN*, *Sim* is the similarity between the attribute value of each member (*u*, *v*) of the pair at time T_i. The calculations of $ETSA^i$ and similarity are given in Equation. 2 and 3 respectively. It is necessary to explain the fact that the Equation 1 considers only the single attribute of each user to measure the similarity. This attribute can be hobby in Facebook or research area in *DBLP*. We believe that Equation 1 can easily be extended to incorporate multiple attributes of the users to measure the similarity:

$$TSA = \frac{\left(TESA^{i} - Min\left(ETSA\right)\right)}{\left(Max\left(ETSA\right) - Min\left(ETSA\right)\right)} \quad (2)$$

where, $ETSA^i$ is the interaction count at time T_i, whereas, *Min*(*ETSA*) and *Max*(*ETSA*) are the minimum and maximum number of interactions between the pairs. The interaction count is normalized, in order to overcome the situation when the interaction count of a pair is so high that it cannot be compared with other potential rising pair. This situation is illustrated in Fig. 2:

$$Sim\left(VTSA_{u}^{i}, VTSA_{v}^{i}\right) = \begin{cases} 1 & if\left(i = N\right) \\ 0 < x < 1 & if\left(i < N\right) \\ 0 & if\left(i = 0\right) \end{cases} \quad (3)$$

We formally define the function Rising_Pairs as Rising_Pairs (*THN*, Dim_1, Dim_2,..., Dim_n) Return [list of rising pairs]. The function takes as input the *THN*, a list of dimensions in which to find the rising pairs and returns a list of rising pairs in the *THN*.

It may be argued that rising pairs can also be computed by writing multiple SQL queries comprising of various existing aggregation functions. However, we believe that sometimes it is not a straightforward job using existing aggregation functions. Therefore, our function avoids the requirements of strong SQL skills and enables the users to focus on data analysis rather than first deriving the data as the case in (Borzsony *et al*., 2001).

Examples of Rising Pairs

Consider Table 1, it shows the co-authorship pattern of three different pairs of authors of DBLP THN. The column Time Interval shows various time instances when the network snapshots were taken. *ETSA* represents the number of co-authored papers in each time interval and VTSA displays the resemblance of two authors in terms of similarity of research areas. Rising_Value for each time interval is calculated using Equation 1. Total indicates the aggregated values of each pair.

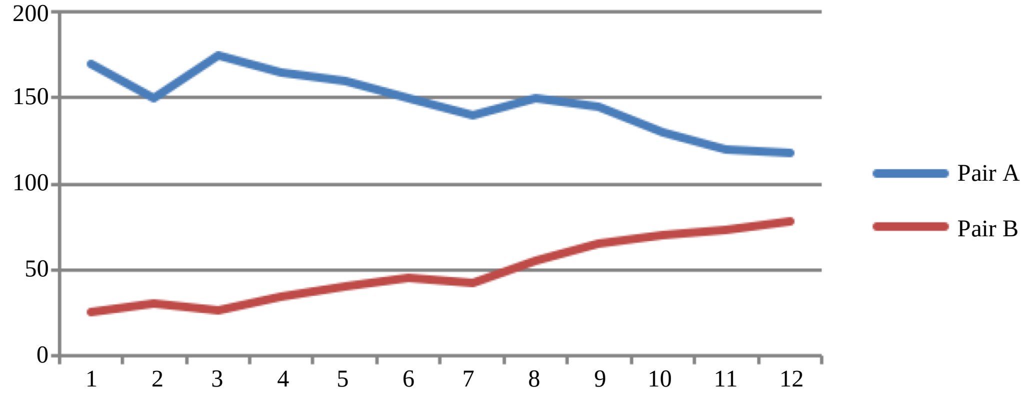

Fig. 2: Sample interactions of pair of vertices A and B. Here *x*-axis denotes count of interactions among a pair of nodes and *y*-axis shows period of 12 months

Table 1: Calculating the Rising_Value (i.e., RValue) from a sample DBLP THN

	Pair 1-2			Pair 3-4			Pair 5-6		
Time interval	ETSA	VTSA	RValue	VTSA	RValue	VTSA	RValue	ETSA	ETSA
T1	7	0.7	0.84	3	0.6	0.74	1	0.0	0
T2	4	0.6	0.68	1	0.1	0.1	1	0.1	0.1
T3	3	0.4	0.46	1	0.1	0.1	1	0.2	0.2
T4	1	0.1	0.1	1	0.3	0.3	1	0.2	0.2
T5	1	0.1	0.1	2	0.3	0.46	2	0.6	0.76
T6	1	0.1	0.1	2	0.5	0.75	3	0.8	1.3
T7	1	0	0	3	0.7	1.7	3	0.9	1.9
Total	18	0.29	2.29	13	1.55	4.15	12	2.8	4.46

We observe that the pair 1-2 has highest number of publications and their similarity is initially high, but as the time goes by, there is a decrease in collaborative work and similarity. The net results show much lower Rising_Value compared to other pairs whose situation was opposite to pair 1-2. There lies the effectiveness of the function Rising_Pairs that it brings forth the pairs whose simple quantitative results are not prominent but there is great potential in them. All the existing aggregation functions are limited in their capabilities to explore such underlying phenomenon and merely reflect the summarized results that do not show the actual situation.

Algorithm

Algorithm 2 presents the pseudo code for identifying the rising pairs in THN. It takes THN as an input along with the dimension(s) and operates iteratively for each pair of vertices across all the snapshots. At line 1, pair having highest aggregated interactions count using its ETSA is identified and is named as $(u, v)_H$. Line 2 determines its Rising_Value. From lines 3-13, the entire THN is traversed. Lines 4-13 operate on each pair of in THN. At line 5 the Rising_Value of current pair is determined. Lines 6 and 9 compares the rising value of $(u, v)_H$ and that of current pair. If rising value of current pair is high or comparable to that of $(u, v)_H$, it is declared as the rising pair. By comparable, we mean that although it is less than that of $(u, v)_H$ but increasing with the passage of time and is expected to be significant later on.

Performance Concerns

The aggregation functions are categorized into distributive, algebraic and holistic functions. The distributive and algebraic functions rely on intermediate results, whereas, the holistic functions need to access the base level tuples for their computation. So, these kind of functions are difficult to optimize (Lenz and Thalheim, 2001). Unfortunately, the proposed function is closer to holistic category. It iterates through each snapshot of network and incrementally calculates the Rising_Value, which is computationally expensive when number of snapshots are really high in number.

Algorithm 2: Rising Pairs

```
  Input: Timed Historical Network, List of Dimensions
  Output: List of rising/emerging pairs in the network
1 Identify the pair $(u, v) \in V$ having highest aggregated
  value of ETSA, known as $(u, v)_H$, in given dimensions;
2 Calculate Rising_Value for $(u, v)_H$ by Equation 1;
3 while THN. End of File do
```

```
  for each pair $(u, v) \in V$ and $(u, v) \neq (u, v)_H$ do
      Calculate Rising_Value for $(u, v)$ by Equation 1;
      if Rising_Value of $(u, v)$ ¿ Rising_Value of $(u, v)_H$ then
          declare $(u, v)$ as rising pair;
      end
      else if Rising_Value of $(u, v)$ and Rising_Value of $(u, v)_H$ comparable then
          declare $(u, v)$ as rising pair;
      end
  end
end
```

Experimental Study

In order to verify the effectiveness of proposed function, we utilized it on a variety of real-world collaborative and social networks. The reason to select the different kinds of dataset is to show the usefulness of the proposed function in different domains. The emphasis is to highlight the fact that such phenomenon is prevailing in different environments. There is a need to explore new functions, like Rising_Pairs, whose focus is on discovering hidden information.

Datasets

In this section, we explain the social networks datasets used for validating the Rising_Pairs.

Facebook

We used the Facebook dataset containing the wall postings between a number of users at various time intervals, made publicly available by (Viswanath *et al*., 2009). We found the maximum number of wall posts in the year 2008, so we extracted the data for this year only from the dataset. The dataset was then divided into 12 snapshots of one month each in order to arrange it in the form of THN.

DBLP

The PROXIMITY DBLP[1] database is based on data from the DBLP Computer Science Bibliography with additional preparation performed by the Knowledge Discovery Laboratory, University of Massachusetts Amherst.

We divided the dataset into seven partitions of five years each from i.e., [1998-2000), [2000, 2002), [2002, 2004), [2004, 2006), [2006, 2008), [2008, 2010) and [2010, 2012]. Each interaction between any two authors is attached with an attribute, holding the number of co-authored papers in a given time interval. Each author is attached with the attribute specifying his/her research area. The research area is figured out using the conference of publications. Since, all the conferences fall into various research categories, so they are combined according to their Focus of Research (FoR). This creates the dimensional hierarchy for the authors as displayed in Fig. 3. We considered only the first two authors of each paper for co-authorship network. The categorization of conferences into FoRs is achieved using the resources provided by "The Computing Research and Education Association of Australasia (CORE)"[2]. The observed FoRs are given in Table 2. During data pre-processing, it was found that the name of some of the conferences, existing in DBLP dataset, is missing in the list maintained by CORE.

GeoLife Trajectory Dataset

This dataset was collected during (Microsoft Research Asia) Geolife project of 182 users in a period of over three years (April 2007-August 2012) and was downloaded from Microsoft research webpage[3]. The dataset contains 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48,000+ hours. These trajectories were recorded by different GPS loggers and GPS-phones and have a variety of sampling rates. Ninety-one percent of the trajectories are logged in a dense representation, e.g., every 1 5 seconds or every 5 10 met per point. A GPS trajectory in this dataset is represented by a sequence of time-stamped points, each of which contains the information about latitude, longitude, date and time and also a label of used transportation mode. Available transportation modes are Walk, Bus, Bike, Car, Taxi and subway. Dataset is distributed into 6 equal time intervals, each containing data of 6 months i.e., [Jan,2007-Jun,2007], [Jun,2007-Dec,2007], [Jan,2008-Jun,2008), [Jun,2008-Dec,2008], [Jan,2009-Jun,2009) and [Jun,2009-Dec,2009]. Moving objects are divided into 11 groups on the basis of their similar locations.

Table 2: Focus of research for conferences

1.	Information and Computing Sciences
2.	Computer Software
3.	Computation Theory and Mathematics
4.	Data Format
5.	Other Information and Computing Sciences
6.	Distributed Computing
7.	Artificial Intelligence and Image Processing
8.	Information Systems
9.	Design Practice and Management

[1] http://kdl.cs.umass.edu/data/dblp/dblp-info.html

[2] http://www.core.edu.au/team

[3] http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

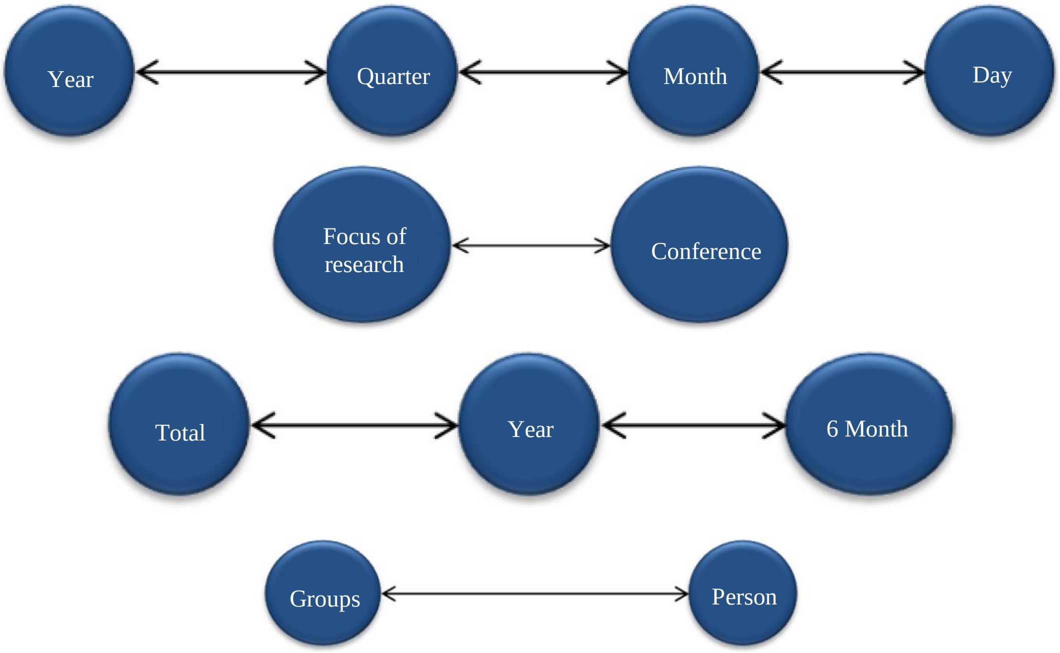

Fig. 3: The dimensional hierarchy for time, FoR and person dimensions

Experimental Results

In this section, we discuss various aspects of results generated from three different datasets such as Facebook, DBLP and GeoLife Trajectory Data respectively.

Facebook

As OLAP is good at multidimensional data analysis, so vertices must be attached with multiple attributes/dimensions. Unfortunately, the available dataset does not have any user's attribute due to privacy concerns. In order to overcome this issue, we synthetically generated the similarity between users by declaring two users are more similar if there are high number of wall posts between them. We set a constant similarity value for users according to their number of wall posts; this value was added by a random number between 0 and 1 in order to remove any kind of overweight to any pair. The edge between each pair is attached with the time dimension to aggregate the count of wall postings at various hierarchical levels. The dimensional hierarchy is displayed in Fig. 3.

The roll-up operation is formulated by analyzing the dataset at the "Quarter" level of the dimensional hierarchy. Figure 4 displays the total wall posts between each other of various pairs. The pair "9137-41668" has the highest aggregated count, so can be termed as the most strongly connected. However, when their Rising_Value is considered, we get different observations in Fig. 5. We find that the recent interactions of this pair are very low in recent times while that of "2286-2277" is quite high, though their aggregated count is lower. Hence, it is unfair to declare them as the strongly connected pairs.

DBLP

Now we demonstrate the application of *Rising_pairs* on DBLP dataset. The function identifies different pairs of authors at various levels of dimensional hierarchy displayed in Fig. 3. There are different criteria at various levels of hierarchy to calculate the similarity of two authors. At FoR level, two authors are similar if they share the same research area; while they are similar when they appear together in the same conference at the conference level. The roll-up operation is formulated in such a way that the rising pairs are discovered at FoRs level. So, we point the rising pairs at the highest level of the dimensional hierarchy.

Figure 6 indicates the total co-authored papers of various authors. We observe that the pair "Irith Pomeranz-Sudhakar M. Reddy" is on top with respect to having highest count but their recent interaction and similarity is lesser to that of "Shusaku Tsumoto-Shoji Hirano". This fact is also illustrated in Fig. 7 and 8 where the roll-up and drill down operation are displayed, showing the relationship at FoR and conference level. The recent relationship is on the decreasing stages while that of "Shusaku Tsumoto-Shoji Hirano" is better and of other pairs too.

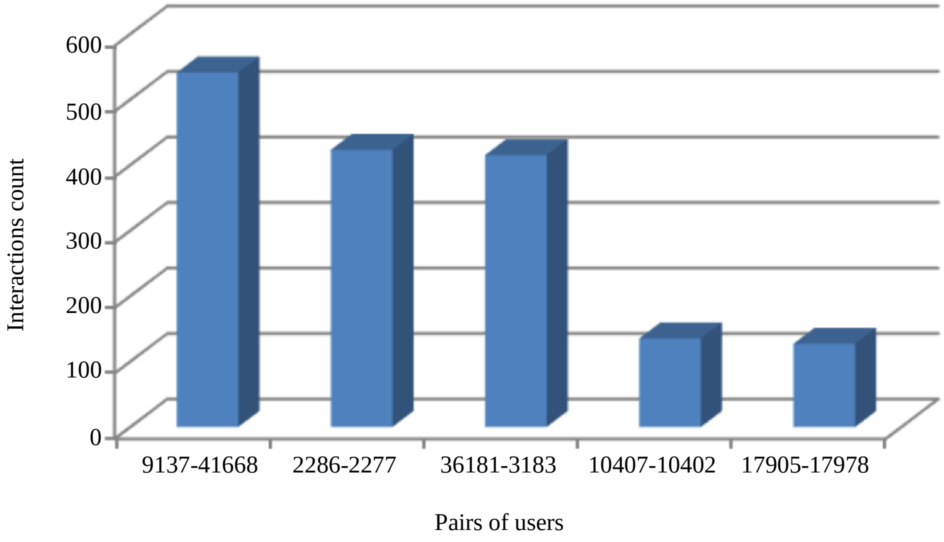

Fig. 4: Total count of interactions from Facebook in 2008

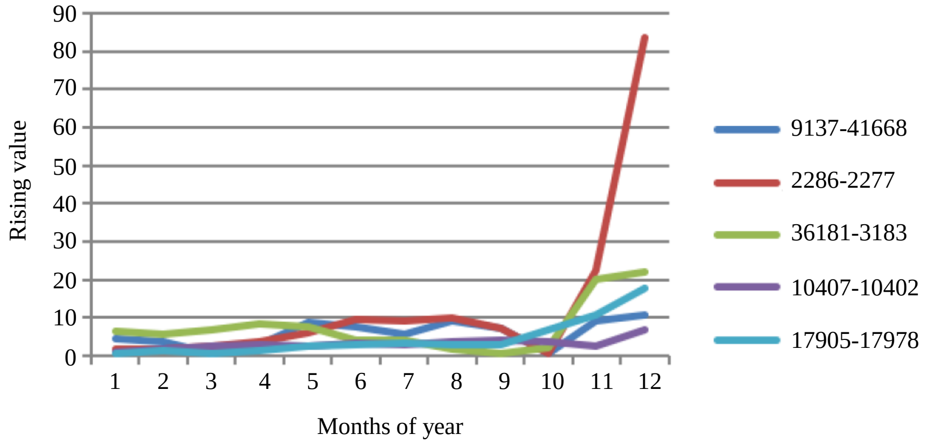

Fig. 5: Rising_Pairs from Facebook in 2008

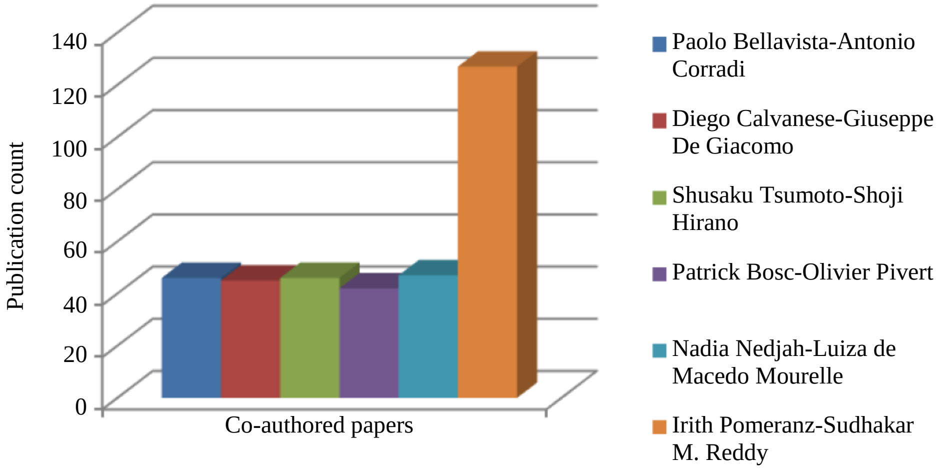

Fig. 6: Author’s interactions

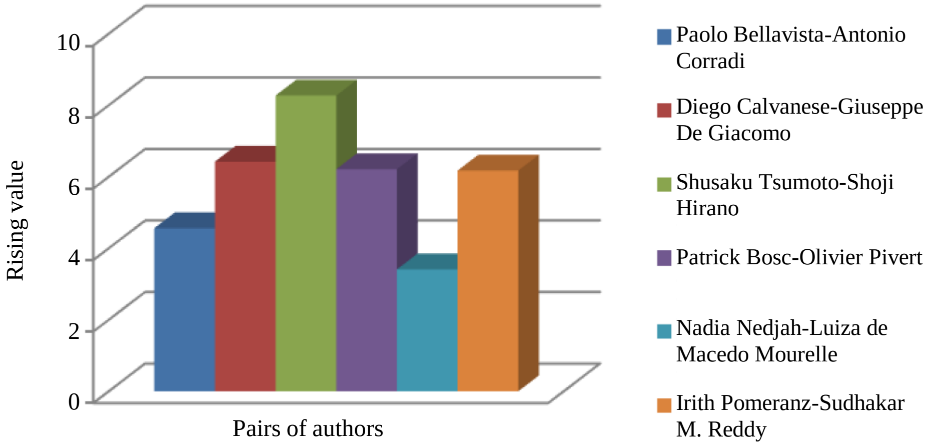

Fig. 7: Rising_Pairs at FoR level

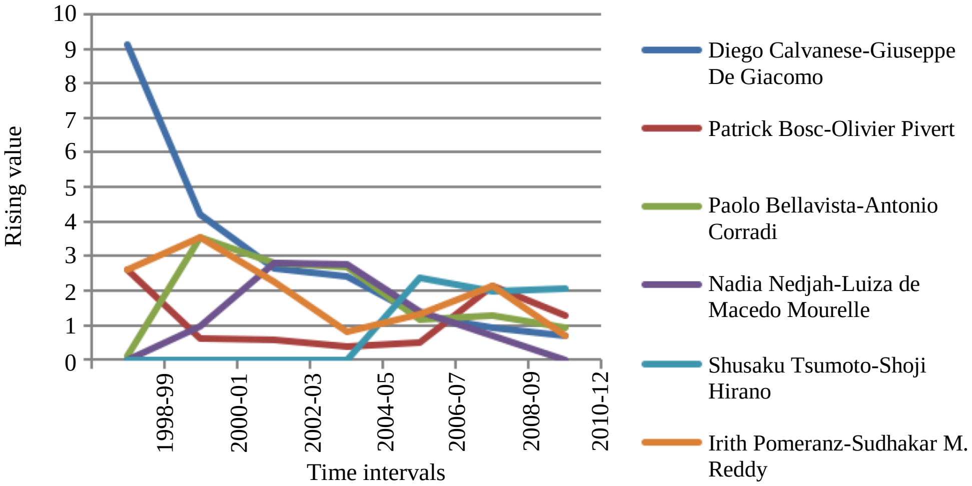

Fig. 8: Rising_Pairs at conference level

GeoLife Trajectory

The dataset is divided into hierarchies based on time intervals and person groups respectively. Figure 3 contains the dimensional hierarchy for time and persons respectively. Attribute considered for each transport mode is its effect in densely populated areas. For example, in case of most inhabited zones using public transport as compared to private is better and can put a positive impact on the environment. Weights are assigned to each transport mode based on its influence on the environment in the following order Walk, Subway, Bike Bus, Taxi, Car. Maximum weight is acquired by Walk and minimum is assigned to the Car. Similarly, each person group is also assigned weight on the basis of their activeness toward solving densely populated environment. Based on our assumption, weights are assigned in the descending order from Group 1 to Group 11. While applying the similarity measure in Rising_Pairs if an object from an active group (having high attribute value) interacts with more high value transport mode an extra weight is assigned to this pair.

In roll-up operation, the time dimension is rolled up to 1 year. Link of person with each transportation mode in all of time intervals are evaluated. Group 8 is chosen due to maximum availability of trajectory data in all data intervals. Figure 9 shows the total edge value between moving objects of group 8 and all transportation modes and rising pairs are exposed in Fig. 10. It can be observed from these figures that the total edge weight of bike, walk was greater than bus and subway respectively. But, Rising_Value shows totally different results with subway value greater than walk and bus more than the bike.

During Drill-down operation, the rising pairs are identified the lowest level of hierarchy for the time dimension. It can be seen in Fig. 11 that Rising_Value of pair of person and the subway is greater than that of others. In this process, we do not consider the presence of the count which is too high. The function Rising_Pairs is showing the groups that are influencing in positive way to solve the issue of transportation in densely populated areas.

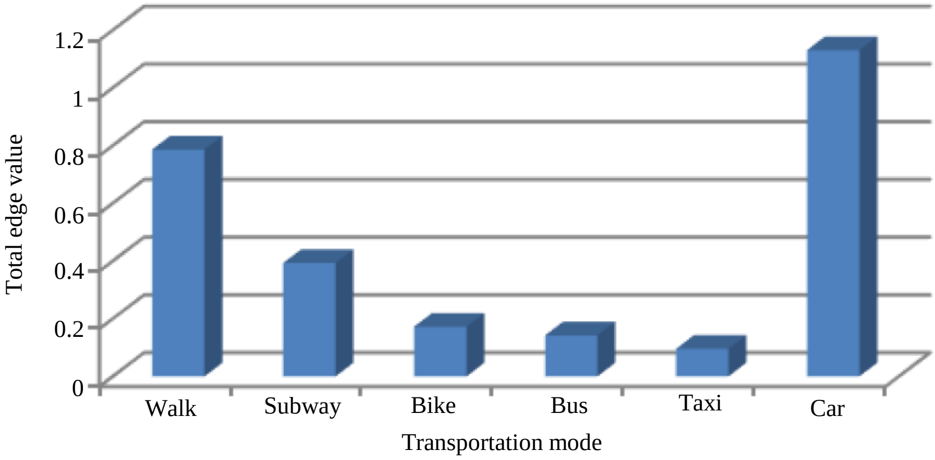

Fig. 9: The frequency of transportation mode on yearly basis

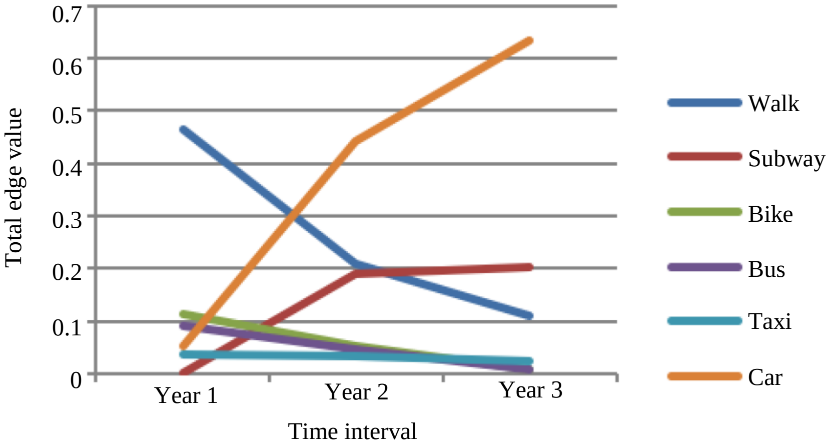

Fig. 10: Rising_Pairs at yearly level

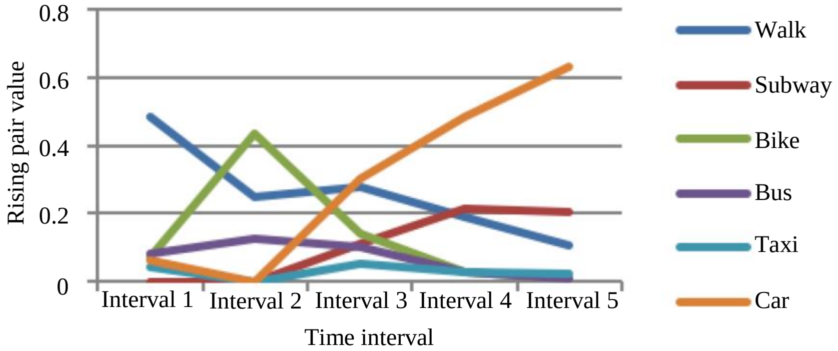

Fig. 11: The frequency of transportation mode on bi-yearly basis

Conclusion

In this study, we proposed an aggregation function, Rising_Pairs, for performing OLAP on collaboration and social networks. The function identifies the strongly connected users in these networks with respect to their recent behavior. It gives more weight to the recent interactions among the people and focuses on users similarities based on their associated attributes. We also proposed the historical and aggregated organization of the underlying networks to perform trend analysis. The significance of historical organization is from the fact that relying on the current status of the network, provides only the limited results. Finally, we validated the usefulness of the function on various real-life datasets like DBLP, Facebook and Microsoft GeoLife project and observed interesting results. Further research is planned to enhance the computational efficiency of the proposed function. We also have a plan to utilize the function towards community detection in social networks. The intention is to see how the communities evolve over the time.

Acknowledgement

This work is done under the grant received by Deanship of research at Islamic University of Madinah (IUM), Saudi Arabia (10th Takamul 1440-1441 AH - Project No. 39/40). We express our thanks to the administration of FCIS and Research deanship at IUM for their support in every aspect of this work. We would like to acknowledge the work done by all the stakeholders of this project, especially, the students of the FAST University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Islamic University of Madinah.

Funding Information

This research was funded by Islamic University of Madinah under the grant of 10th Takamul program of academic year 1440-1441 AH and research project number is 39/40.

Author’s Contributions

Waqas Nawaz: The original draft preparation, project administration and funding acquisition was handled by W. Nawaz.

Kifayat Ullah Khan: The conceptualization of the idea, methodology and formal analysis is from K.U. Khan.

Muhammad Shoaib Siddiqui: The article was reviewed and edited by M.S. Siddiqui.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abazi-Bexheti, L., A. Kadriu and M. Apostolova, 2019. Investigating gender gap in computer science research community. Proceedings of the 42th International Convention on Information and Communication Technology, Electronics and Microelectronics, May 20-24, IEEE Xplore Press, Opatija, Croatia, pp: 737-741. DOI: 10.23919/MIPRO.2019.8757190

Baars, H. and H.G. Kemper, 2008. Management support with structured and unstructured data-an integrated business intelligence framework. Inform. Syst. Manage., 25: 132-148. DOI: 10.1080/10580530801941058

Babskova, A., P. Drazdilova, J. Martinovic, V. Svaton and V. Snasel, 2013. Evolution of co-authors communities formed by terms on DBLP. In DATESO.

Bernabei, C., F. Guerra and R. Trillo-Lado, 2015. Keyword search in structured data and network analysis: A preliminary experiment over DBLP. Proceedings of the 10th International Workshop on Semantic and Social Media Adaptation and Personalization, Nov. 5-6, IEEE Xplore Press, Trento, Italy, pp: 1-6. DOI: 10.1109/SMAP.2015.7370089

Biryukov, M. and C. Dong, 2010. Analysis of computer science communities based on DBLP. Proceedings of the International Conference on Theory and Practice of Digital Libraries, (PDL’ 10), Springer, pp: 228-235. DOI: 10.1007/978-3-642-15464-5_24

Bleco, D. and Y. Kotidis, 2019. Using entropy metrics for pruning very large graph cubes. Inform. Syst., 81: 49-62. DOI: 10.1016/j.is.2018.11.007

Borzsony, S., D. Kossmann and K. Stocker, 2001. The skyline operator. Proceedings of the 17th international conference on data engineering, Apr. 2-6, IEEE Xplore Press, Heidelberg, Germany, pp: 421-430. DOI: 10.1109/ICDE.2001.914855

Burch, M., D. Pompe and D. Weiskopf, 2015. An analysis and visualization tool for DBLP data. Proceedings of the 19th International Conference on Information Visualisation, Jul. 22-24, IEEE Xplore Press, Barcelona, Spain, pp: 163-170. DOI: 10.1109/iV.2015.38

Cabot, J., J.L.C. Izquierdo and V. Cosentino, 2018. Are *cs* conferences (too) closed communities? Commun. ACM, 61: 32-34. DOI: 10.1145/3209580

Calvo-Armengol, A. and M.O. Jackson, 2004. The effects of social networks on employment and inequality. Am. Econ. Rev., 94: 426-454. DOI: 10.1257/0002828041464542

Cesarini, M., F. Mercorio, M. Mezzanzanica, V. Moscato and A. Picariello, 2018. Graphdblp released: Querying the computer scientists network as a graph.

Chen, C., X. Yan, F. Zhu, J. Han and S.Y. Philip, 2008. Graph OLAP: Towards online analytical processing on graphs. Proceedings of the 8th International Conference on Data Mining, Dec. 15-19, IEEE Xplore Press, Pisa, Italy, pp: 103-112. DOI: 10.1109/ICDM.2008.30

Chui, C.K., B. Kao, E. Lo and D. Cheung, 2010. Solap: An OLAP system for analyzing sequence data. Proceedings of the International Conference on Management of Data, (CMD’ 10), ACM, pp: 1131-1134. DOI: 10.1145/1807167.1807299

Fathalla, S., S. Vahdati, S. Auer and C. Lange, 2018. Metadata analysis of scholarly events of computer science, physics, engineering and mathematics. Proceedings of the International Conference on Theory and Practice of Digital Libraries, (PDT' 18), Springer, pp: 116-128. DOI: 10.1007/978-3-030-00066-0_10

Freeman, L.C., 1977. A set of measures of centrality based on between ness. Sociometry.

George, B. and S. Shekhar, 2008. Time-aggregated graphs for modeling spatiotemporal networks. J. Data Sema., 11: 191-212. DOI: 10.1007/978-3-540-92148-6_7

Ghrab, A., O. Romero, S. Skhiri and E. Zimanyi, 2020. Topograph: An end-to-end framework to build and analyze graph cubes. Inform. Syst. Frontiers.

Girvan, M. and M.E. Newman, 2002. Community structure in social and biological networks. Proc. National Acad. Sci., 99: 7821-7826. DOI: 0.1073/pnas.122653799

Han, J., Y. Chen, G. Dong, J. Pei and B.W. Wah *et al*., 2005. Stream cube: An architecture for multi-dimensional analysis of data streams. Distributed Parallel Databases, 18: 173-197. DOI: 10.1007/s10619-005-3296-1

Herrmannova, D., 2018. Mining scholarly publications for research evaluation. PhD Thesis, The Open University.

Huberman, B.A., D.M. Romero and F. Wu, 2008. Social networks that matter: Twitter under the microscope.

Kaufman, L. and P.J. Rousseeuw, 2009. Finding Groups in Data: An Introduction to cluster Analysis. 99th Edn., John Wiley and Sons, ISBN-10: 0470317485, pp: 342.

Keselman, L., 2019. Venue analytics: A simple alternative to citation-based metrics. Proceedings of the Joint Conference on Digital Libraries, Jun. 2-6, IEEE Xplore Press, Champaign, IL, USA, pp: 315-324. DOI: 10.1109/JCDL.2019.00052

Khan, K.U., W. Nawaz, M.A. Saleem, Y.K. Lee and S. Lee, 2012. Rising_Pairs: An OLAP aggregation function for social networks. Proceedings of the 4th international conference on Emerging Databases-Technologies, Applications and Theory, (TAT' 12), Seoul, Republic of Korea.

Kim, H., J.W. Yoon and J. Crowcroft, 2012. Network analysis of temporal trends in scholarly research productivity. J. Inform., 6: 97-110. DOI: 10.1016/j.joi.2011.05.006

Kimball, R. and M. Ross, 2011. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 1st Edn., John Wiley and Sons.

Kumar, N., R. Utkoor, B.K. Appareddy and M. Singh, 2017. Generating topics of interests for research communities. Proceedings of the International Conference on Advanced Data Mining and Applications, (DMA' 17), Springer, pp: 488-501. DOI: 10.1007/978-3-319-69179-4_34

Lange, P.G., 2007. Publicly private and privately public: Social networking on Youtube. J. Comput. Med. Commun., 13: 361-380. DOI: 10.1111/j.1083-6101.2007.00400.x

Lawrance, A. and P. Lewis, 1977. An exponential moving-average sequence and point process (ema1). J. Applied Probability, 14: 98-113. DOI: 10.2307/3213263

LeFevre, K. and E. Terzi, 2010. Grass: Graph structure summarization. Proceedings of the International Conference on Data Mining, (CDM' 10), SIAM, pp: 454-465. DOI: 10.1137/1.9781611972801.40

Lenz, H.J. and B. Thalheim, 2001. OLAP databases and aggregation functions. Proceedings of the 13th International Conference on Scientific and Statistical Database Management, Jul. 18-20, IEEE Xplore Press, Fairfax, VA, USA, pp: 91-100. DOI: 10.1109/SSDM.2001.938542

Li Zhang, L., R. Adamic and E. Lukose, 2004. Implicit structure and the dynamics of blogspace. Commun. ACM: CACMa PUBL Assoc. Comput. Mach., 47: 35-39. DOI: 10.1145/1035134.1035162

Li, Y. and H. Shen, 2011. Anonymizing graphs against weight-based attacks with community preservation. J. Comput. Sci. Eng., 5: 197-209. DOI: 10.5626/JCSE.2011.5.3.197

Liu, Y., J. Li and H. Gao, 2008. Summarizing graph patterns. Proceedings of the 24th International Conference on Data Engineering, Apr. 7-12, IEEE Xplore Press, Cancun, Mexico, pp: 903-912. DOI: 10.1109/ICDE.2008.4497499

Lo, E., B. Kao, W.S. Ho, S.D. Lee and C.K. Chui *et al*., 2008. OLAP on sequence data. Proceedings of the International Conference on Management of Data, (CMD' 08), ACM, pp: 649-660. DOI: 10.1145/1376616.1376682

Mercorio, F., M. Mezzanzanica, V. Moscato, A. Picariello and G. Sperlı, 2019a. A Tool for Researchers: Querying big scholarly data through graph databases. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, (KDD' 19).

Mercorio, F., M. Mezzanzanica, V. Moscato, A. Picariello and G. Sperli, 2019b. DICO: A Graph-DB framework for community detection on big scholarly data. Trans. Emer. Topics Comput.

Mezzanzanica, M., F. Mercorio, M. Cesarini, V. Moscato and A. Picariello, 2018. Graphdblp: A system for analysing networks of computer scientists through graph databases. Multi. Tools Applic., 77: 18657-18688. DOI: 10.1007/s11042-017-5503-2

Moreira, C., P. Calado and B. Martins, 2015. Learning to rank academic experts in the DBLP dataset. Exp. Syst., 32: 477-493. DOI: 10.1111/exsy.12062

Navlakha, S., R. Rastogi and N. Shrivastava, 2008. Graph summarization with bounded error. Proceedings of the International Conference on Management of Data, (CMD' 08), ACM, pp: 419-432.

Orr, L. and J. Ortiz, 2013. Clustering with the DBLP bibliography to measure external impact of a computer science research area.

Pflugrad, A., 2017. Developing an automated bibliometric analysis system for finding rare disease experts. PhD Thesis, Universitat Ulm.

Qu, Q., F. Zhu, X. Yan, J. Han and S.Y. Philip *et al.*, 2011. Efficient topological OLAP on information networks. Proceedings of the International Conference on Database Systems for Advanced Applications, (SAA' 11), Springer, pp: 389-403. DOI: 10.1007/978-3-642-20149-3_29

Queiroz-Sousa, P.O. and A.C. Salgado, 2019. A review on OLAP technologies applied to information networks. ACM Trans. Knowledge Dis. Data, 14: 1-25. DOI: 10.1145/3370912

Ravat, F., O. Teste and R. Tournier, 2007. Olap aggregation function for textual data warehouse. ICEIS.

Ravat, F., O. Teste, R. Tournier and G. Zurfluh, 2008. Top keyword: An aggregation function for textual document OLAP. Proceedings of the International Conference on Data Warehousing and Knowledge Discovery, (WKD' 08), Springer, pp: 55-64. DOI: 10.1007/978-3-540-85836-2_6

Ryu, B.K., 2020. The demise of single-authored publications in computer science: A citation network analysis.

Seo, Y.J., H.M. Cho and S. Huh, 2020. Changes in bibliographic information associated with Korean scientific journals from 2011 to 2019. Sci. Editing, 7: 11-15. DOI: 10.6087/kcse.184

Shirani-Mehr, H., F.B. Kashani and C. Shahabi, 2012. Efficient reachability query evaluation in large spatiotemporal contact datasets.

Song, M., G.E. Heo and S.Y. Kim, 2014. Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in DBLP. Scientometrics, 101: 397-428. DOI: 10.1007/s11192-014-1246-2

Subramani, M.R. and B. Rajagopalan, 2003. Knowledge-sharing and influence in online social networks via viral marketing. Commun. ACM, 46: 300-307. DOI: 10.1145/953460.953514

Swaraj, K. and D. Manjula, 2016. A fast approach to identify trending articles in hot topics from xml based big bibliographic datasets. Cluster Comput., 19: 837-848. DOI: 10.1007/s10586-016-0561-1

Tang, J., J. Sun, C. Wang and Z. Yang, 2009. Social influence analysis in large-scale networks. Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining, (DDM' 09), ACM, pp: 807-816. DOI: 10.1145/1557019.1557108

Tian, Y., R.A. Hankins and J.M. Patel, 2008. Efficient aggregation for graph summarization. Proceedings of the International Conference on Management of Data, (CMD' 08), ACM, pp: 567-580. DOI: 10.1145/1376616.1376675

Ueda, A.H., B.R. de Araujo Neto, E.D.S.E Silva, M. Dias and N. Ziviani, 2017. Reputation in computer science on a per subarea basis. Technical Report.

Viswanath, B., A. Mislove, M. Cha and K.P. Gummadi, 2009. On the evolution of user interaction in facebook. Proceedings of the 2nd Workshop on Online Social Networks, (OSN' 09), ACM, pp: 37-42. DOI: 10.1145/1592665.1592675

Yang, Z., L. Hong and B.D. Davison, 2013. Academic network analysis: A joint topic modeling approach. Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, (NAM' 13), ACM, pp: 324-333. DOI: 10.1145/2492517.2492524

Zhao, P., X. Li, D. Xin and J. Han, 2011. Graph cube: On warehousing and OLAP multidimensional networks. Proceedings of the 2011 International Conference on Management of Data, (CMD' 11), ACM, pp: 853-864. DOI: 10.1145/1989323.1989413