Transactions of the Indian National Academy of Engineering (2022) 7:365–374
https://doi.org/10.1007/s41403-021-00306-y

ORIGINAL ARTICLE

# Topological Analysis on Multi-scenario Graphs: Applications Toward Discerning Variability in SARS-CoV-2 and Topic Similarity in Research

**Sourav Biswas**[1,2] · **Malay Bhattacharyya**[3] · **Sanghamitra Bandyopadhyay**[1]

Received: 16 May 2021 / Accepted: 20 November 2021 / Published online: 4 February 2022


**Abstract**
A network is often an obvious choice for modeling real-life interconnected systems, where the nodes represent interacting objects and the edges represent their associations. There has been immense progress in complex network analysis with methods and tools that can provide important insights into the respective scenario. In the advancement of information technology and globalization, the amount of data is increasing day by day, and it is indeed incomprehensible without the help of network science. This work highlights how we can model multiple interaction scenarios under a single umbrella to uncover novel insights. We show that a varying scenario gets reflected by the change of topological patterns in interaction networks. We construct multi-scenario graphs, a novel framework proposed by us, from real-life environments followed by topological analysis. We focus on two different application areas: analyzing geographical variations in SARS-CoV-2 and studying topic similarity in citation patterns.

**Keywords** Topological Analysis · Multi-scenario Graphs · Graphlet and Graphlet Degree Distribution · SARS-CoV-2 · Citation Network · Topic Similarity Network

## Introduction

Network analysis is one of the essential tasks for an application domain that deals with interconnected systems. This is mainly for two reasons. First of all, the network provides an excellent visual representation to realize the environment and build intuitive ideas Pavlopoulos et al. (2008), So et al. (2020), Jennifer and Chen (2005). Moreover, the complexity of an interactive system lies in the collaborative contribution of many parts of the system; hardly, a single element can do a task Booher and Innes (2002), Lorts et al. (2020). Therefore, a network model is the best way to reflect various approaches to symbolize the interactions between real-life objects.

On many occasions, we deal with environments that encompass different sets of homogeneous objects connected in a similar context. In biology, this could be thought of as networks between different genes active in different tissue types or country-specific networks between viral strains (say, the strains of SARS-CoV-2). With a completely different perspective, this could also be considered as networks of collaborating researchers across different years, country-specific networks of collaborating researchers, etc. We can model these scenarios with a novel concept of multi-scenario graphs. This will consider the networks (between similar kinds of objects) evolving from multiple but similar scenarios.

In this paper, we propose the novel concept of analyzing multi-scenario graphs as a generic approach of studying multiple complex networks. For demonstrating its genericness, we employed it for two completely different applications. One is understanding the variability of SARS-CoV-2 (the virus attributed to COVID-19) strains across multiple countries, and the other is unearthing topic similarity in research publication networks. The motivations behind

✉ Sanghamitra Bandyopadhyay
sanghami@isical.ac.in

Sourav Biswas
sourav8051_r@isical.ac.in; sourav8051@gmail.com

Malay Bhattacharyya
malaybhattacharyya@isical.ac.in

[1] Indian Statistical Institute, Kolkata, India

[2] University of Calcutta, Kolkata, India

[3] Machine Intelligence Unit Centre for Artificial Intelligence and Machine Learning Technology Innovation Hub on Data Science, Big Data Analytics, and Data Curation, Indian Statistical Institute, Kolkata, India

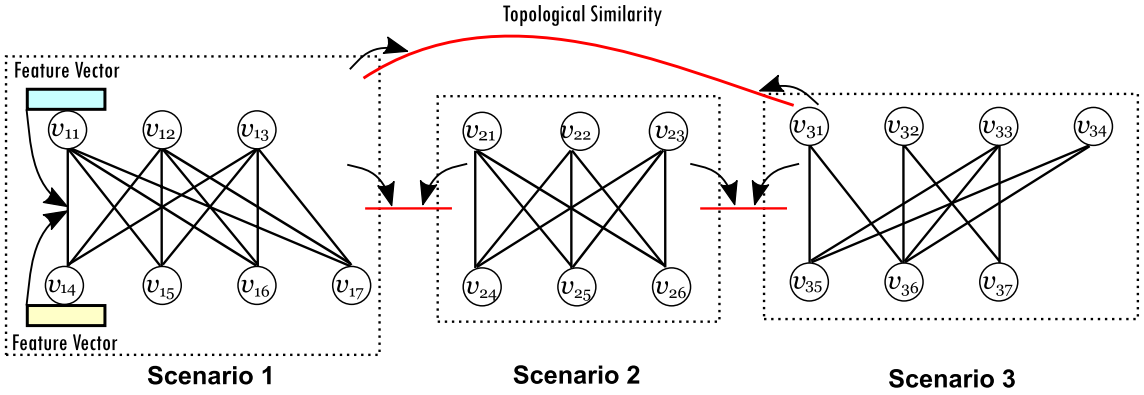



**Fig. 1** The approach of topological analysis on multi-scenario graphs is shown. Different scenarios are depicted with separate graphs in which the nodes are connected in a similar context. Based on topological similarity, further edges (with weights) are defined between the said scenarios. This leads to a multi-scenario graph. Edges in the individual scenario graphs and in the multi-scenario graph are colored in black and red, respectively

these applications are described hereunder. The Sect. 2 contains basic terminologies and Sect. 3 describes some related works. After that, in the Sect. 4, we state the methods as a generic version followed by the Sect. 5, where we applied this method in two different applications.

In the first application, we consider hundreds of viral samples causing COVID-19 that are being sequenced all over the world. A $k$-mer distribution of a sequence is the count of nucleotides of size $k$ found in it. The value of $k$ should be compromised between computational complexity and the characteristics of the sequence. We found the $k$-mer distribution of the virus sequences collected from different geographical regions and built a $k$-mer network by embedding that vector into a node. We found differences among countries and continents, which supports the hypothesis of fast mutation of this virus.

In the other application, we work on citation networks that contain paper sources (denoting nodes) linked by co-citation relationships (representing edges) Cawkell (1971), Hummon and Dereian (1989). An author name and many other details, like his/her age, country, research domain, university, date of publication and citation, h-index, the total number of papers, the total number of citations, etc., are important. It may be interesting to find some patterns of these networks across the country, decade, different domains etc., which can get a thorough insight into the citation information. Using a multi-scenario graphical approach, we can find the real associative or engagement between people. Each person has his/her unique features or characteristics, which can be taken as a vector and then mapping that vector as a node in the embedded network seems justified.

## Basic Terminologies and Definitions

Let us first introduce the following necessary definitions.

**Definition 1** (*Scenario Graph*) A scenario graph, $G = (V, E)$, is defined over a set of vertices $V = \{v_1, v_2, \ldots, v_{|V|}\}$ of similar type and a set of edges $E : (v_i, v_j)$ $(v_i \neq v_j, \forall v_i, v_j \in V)$ defined for a particular scenario.

The order and size of a scenario graph define the number of nodes and edges it comprises. Let us assume $|S|$ denotes the cardinality (number of elements) of a set $S$. The other notations have their usual meaning, unless specified otherwise. We now formally define a multi-scenario graph as follows.

**Definition 2** (*Multi-scenario Graph*) A multi-scenario graph, $MSG = (V_G, E_G)$, is defined over a set of scenario graphs $V_G = \{G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_n = (V_n, E_n)\}$ (such that $V_1 \cap V_2 \cap \ldots \cap V_n = \emptyset$) and $E_i$s are defined in a similar context) and a set of edges $E_G : (G_i, G_j)$ $(G_i \neq G_j, \forall G_i, G_j \in V_G)$.

Note that, a multi-scenario graph is constructed from multiple graphs that evolve from different scenarios. These scenarios could be spatial, temporal or anything else. However interestingly, the interactions in the said scenarios are defined between node pairs in a similar context. In this work, we aim to construct such multi-scenario

graphs (see Fig. 1) in different cases and perform topological analysis for exploratory study.

## Related Work

Topological analysis of graphs is an important aspect of many different applications Malay and Sanghamitra (2010), Ha et al. (2020), Dongsheng et al. (2021). Finding similarity or dissimilarity between two or more graphs is a hard task, as mentioned in Pržulj (2007), and the author proposed a similarity method based on graphlet degree distributions. In Biswas et al. (2018), an idea has been proposed to explore the similarities of multiple graphs using normalized weighted graphlet frequency distribution. Another approach is called graph embedding, which transforms a graph's properties to a vector that captures the topology, relationship among the vertices, and much other relevant information such as its degree distribution, motifs, anti-motifs, frequent graphlets, special subgraphs, cliques, etc. There has been extensive literature and different approaches of topological graph analysis like planar embedding István (1948), Tutte embedding Tutte (1963), progressive embedding Shen et al. (2019). In some papers, a different idea is proposed for multiple graph topological analysis by dimensionality reduction Yan et al. (2005), Yan et al. (2006), which integrates multiple state-of-the-art supervised or unsupervised algorithms within a common framework. The authors proposed a new method called "Marginal Fisher Analysis" and used it in face recognition applications. In Biswas and Bandyopadhyay (2020), the authors showed a multi-scenario embedding approach for graph analysis. Very recently, progressive embedding has been proposed for complex graph analysis Shen et al. (2019).

However, there is no such approach that encounters the issue of topological insights on multi-scenario graphs. This paper proposes a novel method to explore the said problem and its application in wide domains.

## Method

The proposed method can be divided into three major subtasks. These are (I) Creating an ensemble of networks by some state-of-the-art methods from multiple datasets within a similar domain, (II) Extracting some features from the generated multi-scenario graphs' topology, and (III) Applying a suitable graph analysis or graph comparison methods to get an efficient insight into the graphs and the given original dataset. For the third step we have used a topological metric called 'graphlet degree distribution', which was proposed by Prvzulj et al. in an earlier work Pržulj (2007). It is an extension of the traditional degree distribution. Here, 'degree' of a particular graphlet is the number of nodes touching anyone of its nodes.

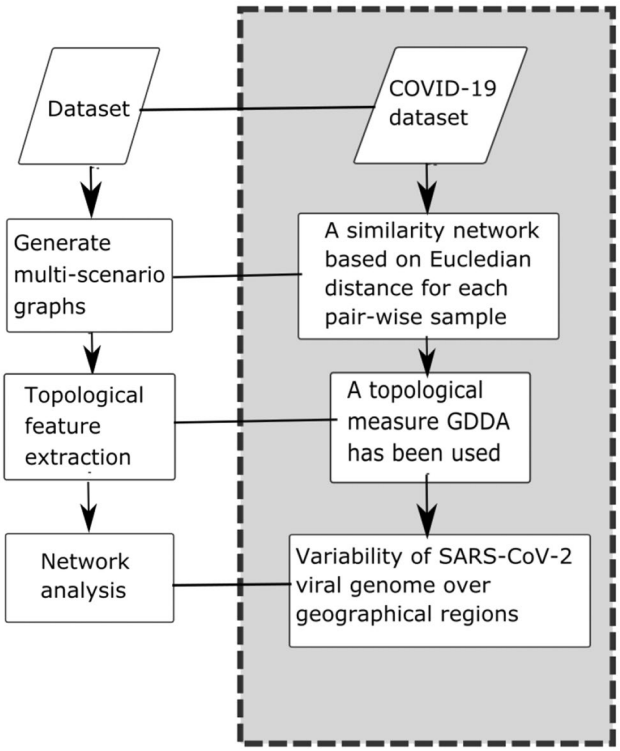



**Fig. 2** Flowchart of the proposed method and corresponding steps done with COVID-19 data as an example (contained in the gray-colored dashed box)

The proposed approach is formally presented below and the flowchart of the following steps is shown in Fig. 2.

1. **Generate Networks**: Given multiple datasets over a particular problem or domain, construct an ensemble of multi-scenario networks/graphs. For example, in our first application with COVID-19 data, we generated *k*-mer distribution from networks for different countries and time spans. Thus a set of networks are generated, but each with different but similar objects, in our case, different samples of the same virus. A relationship can be taken as edges in the graphs. Here for our first application, we choose the 'Euclidean distance', and for the second application, we used the 'Jaccard' Distance'.
2. **Topological Feature Extraction from the Generated Multi-scenario Graphs**: Extract topological features from the generated graphs. A graph embedding approach is confined within a single graph; here, it is extended to the multiple graphs which are generated earlier. Since the graphs are of different sizes and consist of different objects, we need to extract features that are normalized and independent.

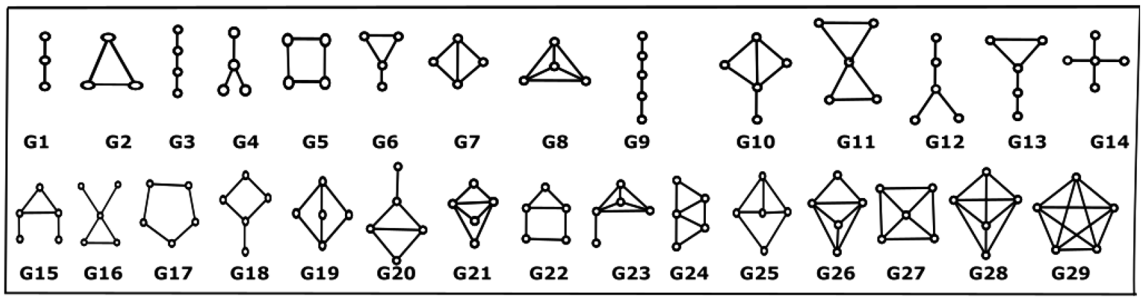



**Fig. 3** All possible 29 non-isomorphic non-directed graphlets or small graphs of order up to 5. The 'graphlet degree distribution' ('Graphlet Degree distribution' is the number of nodes touching each of these 29 graphlets) is used for comparing similarity between two or more non-directed graphs. This measure is known as GDDA

3. **Network Analysis**: One of the topological features of a graph is graphlets degree distribution. A graphlet is a small connected non-induced subgraph of the original network. The graphlet degree distribution is defined as the number of nodes 'touches' node of size-*k* graphlets. A measure using these graphlets' degree distribution, which is known as 'graphlet degree distribution agreement' (GDDA), is employed Pržulj (2007), Shi and Shen (2019). The degree distributions of 29 graphlets (as shown in Fig. 3) are used in this measure. Due to computational complexity, the size of the graphlets is limited to five.

   Note that any other network analysis methods could be done in this step.

## Empirical Analysis

In this study, we mainly aim to construct multi-scenario graphs corresponding to two different application areas, namely the analysis of regional viral strains and citation analysis of research papers. These multi-scenario graphs combine different scenarios, depicted as graphs constructed in a particular context (refer to Fig. 1). To be precise, we demonstrate the usefulness of multi-scenario graph analysis toward analyzing (i) geographical variations in SARS-CoV-2 viruses and (ii) comparing citation networks with topic similarity networks of a well-known KDD Cup 2003 dataset. These are discussed hereunder.

### Application 1: Variability of SARS-CoV-2 Viral Genome over Geographical Regions

COVID-19 is a highly contagious lung disease attributed to the newly discovered novel coronavirus SARS-CoV-2 World Health Organization et al. (2020) originating from Wuhan, China. This disease has caused a long-lasting pandemic since the beginning of 2020. Due to its long incubation period and high infectiousness, COVID-19 has quickly spread all over the world. The World Health Organization (WHO) declared the outbreak of disease as a Public Health Emergency of International Concern in January 2020 and a pandemic on March 11, 2020. Since the virus is newly identified World Health Organization et al. (2020), its transmission mechanisms were unknown in the earlier phase Cox et al. (2020), Kakimoto et al. (2020), and drug practices were arbitrary Afshinnekoo et al. (2021), the threat of this pandemic is still not over. Till date, more than two hundred million people have been infected with COVID-19 worldwide. At the beginning, scientists started understanding the disease by sequencing the SARS-CoV-2 genome (with multiple mutated variants) evolving around the world. As soon as the genome is sequenced, we obtain an in-depth comprehension about the variability of this virus that has spread in different countries. It is already known that being an RNA virus Domingo and Holland (1997) SARS-CoV-2 mutates itself Pachetti et al. (2020) and there are already multiple strains found even in a single country Tang et al. (2020), Phan (2020). Understanding the sequence patterns of this virus across different countries (depicted as scenarios) is therefore important. We employ the multi-scenario graph approach to achieve this.

#### Dataset Preparation

We downloaded SARS-CoV-2 genomic sequence data publicly available at the NCBI Virus portal (https://www.ncbi.nlm.nih.gov/labs/virus, accessed on October 28, 2020) that cover multiple countries and continents. We obtained 22,756 genomic sequences of SARS-CoV-2 collected from humans as the host. Though some virus samples were available from other hosts (animals), we restricted ourselves to a single host in this study. As some of the sequences were released in incomplete form, we performed an initial pre-processing by taking only those with a length of more than 29kb. This resulted in the final selection of 21,273 viral genomes. We

**Table 1** The count of sequences released by the countries with the maximum number of SARS-CoV-2 samples sequenced till October 28, 2020

| Countries | # Released SARS-CoV-2 sequences |
|---|---|
| USA | 12,034 |
| Australia | 7160 |
| India | 556 |
| Egypt | 235 |
| China | 107 |
| Total | 20,092 |

grouped these samples according to their countries. The chosen viral genomes belong to sixty countries The Table 1 shows the five countries that have the maximum number of samples in this set.

We observed that the distribution of the number of SARS-CoV-2 samples sequenced was highly imbalanced both country-wise and continent-wise. Australia and the USA (hence, the corresponding continents) both had a huge number of viral samples as compared to the other countries. Since the dataset had a huge imbalance, we under-sampled Australia and the USA down to around 300 samples. Note that the USA and Australia both had huge numbers of samples in comparison with India, China and Egypt. As our goal was to undersample the sequence data only for the USA and Australia to align with the rest, we took the average of the number of samples available for India, China and Egypt. As this particular average count was 300, hence the choice. Fig. 4a, b show the distributions of the samples before and after under-sampling, respectively.

We obtained the *k*-mer distributions within all the viral genome samples collected from the five countries, namely USA, Australia, India, Egypt and China. A *k*-mer distribution is the frequency of *k*-mers (occurrence of the different combinations of nucleotides of size *k*) in a genomic sequence. We calculated the *k*-mer counts in the genomic sequences by taking values of *k* as 1, 2 and 3. Though in protein synthesis, 3-mers (the codons) play the most crucial role, however, in DNA analysis 1-mer and 2-mers are also important as they may bear signatures for promoter regions, methylation patterns, etc. Hence, the choice.

As in general, there might be only four nucleotides (A, C, T and G) in a genomic sequence, there could be $4^k$ number of possible *k*-mers for each *k*. Hence, for $k = 1$ we had 4 *k*-mers (A, C, T, G), for $k = 2$ we had 16 *k*-mers (AA, AC, ..., GT, GG), and for $k = 3$ we had 64 *k*-mers (AAA, AAT, ..., TTT). So, in total for *k* as 1, 2 and 3, we had 84 *k*-mer counts for each sample. Note that, because the genomic lengths were not necessarily the same in different samples, we normalized the *k*-mer counts by dividing the number of *k*-mers in each sample by the total length of the sequence. These constitute a feature vector corresponding to each sample.

## Creating Scenario Graphs

We reckoned genetic variation between the samples (SARS-CoV-2 genomic sequences) obtained for the five countries to generate five different scenario graphs. The distance between samples refer to the distance between the vectors of *k*-mer counts corresponding to the countries. The distance is calculated using the R package *philentropy* Drost (2018). The 'Euclidean distance' is taken as the distance measure. For each pairwise distance matrix (having a dimension $n \times n$, $n$ being the number of samples for each country), the respective median is chosen as a threshold value for converting it to a binary matrix. The distance values less than the cutoff are converted to ones and others to zeros. Now this binary matrix was used as the adjacency matrix to generate the scenario graphs. Note that a scenario depicts a country in this

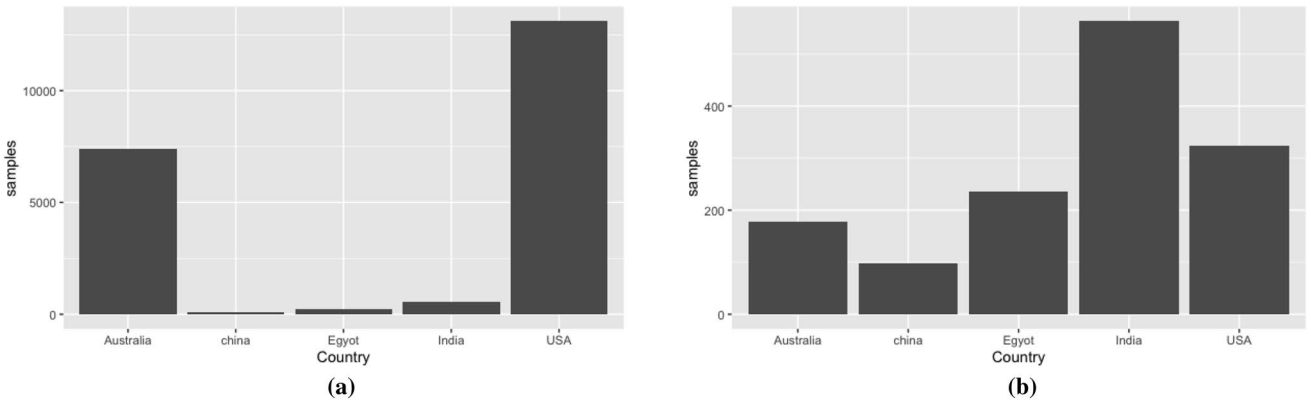



**Fig. 4** **a** The distribution of samples of SARS-CoV-2 genomic sequences among the top five countries which reflects a high imbalance in the dataset. **b** The distribution of samples of SARS-CoV-2 genomic sequences among the top five countries after *k*-fold random under-sampling that balances the distribution

**Table 2** Pairwise similarity scores of *k*-mer networks across the five countries reflecting the geographical mutations in the viral genome till October 28, 2020

| | Australia | India | USA | China | Egypt |
|---|---|---|---|---|---|
| Australia | 1.000000 | 0.240053 | 0.268511 | 0.306087 | 0.280395 |
| India | 0.240053 | 1.000000 | 0.290913 | 0.264563 | 0.289210 |
| USA | 0.268511 | 0.290913 | 1.000000 | 0.309002 | 0.237237 |
| China | 0.306087 | 0.264563 | 0.309002 | 1.000000 | 0.313009 |
| Egypt | 0.280395 | 0.289210 | 0.237237 | 0.313009 | 1.000000 |

It reflects the mutation of the viral genome over multiple countries and over time

case. We finally generated *k*-mer networks corresponding to each country from the said binary matrices.

### Analyzing the Multi-scenario Graph

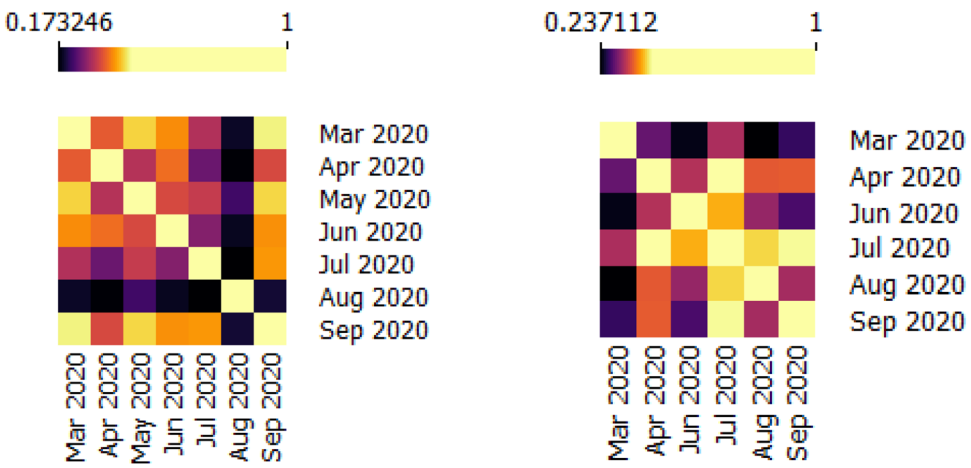
We constructed a multi-scenario graph from the five different scenario graphs based on topological coherence between a pair of graphs. The GDDA is used as a topological similarity measure to find out the coherence between two graphs Pržulj (2007). The `NetworkSim` package in R Shi and Shen (2019) was used to find the pairwise similarity values of the generated *k*-mer networks. The pairwise GDDA scores for the five countries chosen are shown in Table 2.

The pairwise GDDA values between all countries are shown in the Table 2. Since the start of the pandemic and samples collected from different countries were vastly non-identical, it is hard to find any pattern here. For example, China had early genomic sequence data where the pandemic hit in December 2020 with a peak in February 2020. India experienced a later peak in September 2020, and the USA in between. As the SARS-CoV-2 virus mutated very fast with a differential pattern across the globe and the five countries experienced it over a long non-overlapping period of time, it shows no difference. This is well justified by the trends of the pandemic across these countries.

To have a better view of the temporal pattern of the change of the virus, we constructed multi-scenario graphs from the monthly scenarios. We performed pairwise comparison of *k*-mer networks for the two countries Australia and the USA with respect to the monthly data (of sequences collected) between March-December 2020. The respective heatmaps are shown in Fig. 5a, b. It can be seen from the month-wise heatmap of the USA (see Fig. 5a) that there was a sudden change in values after August 2020, which was the time when the second wave hit the country again. On the other side, for Australia (see Fig. 5b) there is a significant difference of k-mer distributions in March 2020 than the others, which possibly indicates a local mutation that triggered the pandemic in Australia.. Thus, analyzing multi-scenario graphs is important to unearth hidden patterns from scenario based interconnected systems.

## Application 2: Topic Similarity in a Research Publication Network with Embedded Keywords

A citation is an acknowledgement to others' work (by including it in the reference) in a publication. This usually highlights the source of information used in a research study. In academics, a study is always related to, motivated by, inspired from or originating from some other study. In this way, the scientific community and human knowledge grow up step by step throughout history. To put in simple words, a citation network or graph consists of some nodes, each




**(a)** USA **(b)** Australia

**Fig. 5** Heatmaps of pairwise GDDA similarity scores between the month-wise scenario graphs for the top two countries, namely the USA (between March-September 2020) and Australia (between March-September 2020 except May due to lack of enough sequence data). These empirical results represent the mutation of the viral genomes in different time-zones of a country. The similarity values (GDDA score) range between [0, 1], with '1' representing the highest similarity

**Table 3** Order and size of (a) citation subnetworks (papers published in a year and the papers who cited those) and (b) topic similarity subnetworks (the papers published in a particular year), respectively for the years 1992–2003

| Year | Order | Size | Year | Order | Size |
|---|---|---|---|---|---|
| (a) | | | (b) | | |
| 1992 | 872 | 1968 | 1992 | 1367 | 46740 |
| 1993 | 1958 | 5299 | 1993 | 2058 | 86277 |
| 1994 | 2775 | 8907 | 1994 | 2377 | 144840 |
| 1995 | 3708 | 12799 | 1995 | 2303 | 114136 |
| 1996 | 4412 | 16489 | 1996 | 2606 | 112936 |
| 1997 | 4847 | 18901 | 1997 | 2673 | 86352 |
| 1998 | 5327 | 22805 | 1998 | 2758 | 122037 |
| 1999 | 3101 | 12423 | 1999 | 2803 | 102497 |
| 2000 | 4139 | 16726 | 2000 | 3126 | 195617 |
| 2001 | 6005 | 48241 | 2001 | 3153 | 127010 |
| 2002 | 2981 | 14636 | 2002 | 3312 | 175793 |
| 2003 | 53 | 67 | 2003 | 1019 | 34743 |

representing a unique article and directed edges between them if a paper $i$ cites another paper $j$.

In a traditional citation network, a node contains only a concept or phrase denoting the name of the article. We argue it captures less information about the different types of citations. As for example, it does not tell whether a citation comes from a common or different domain. Notably, a citation between two papers both from some domain $d_i$ is somewhat different from citations of papers from domain $d_i$ to $d_j$.

### Dataset Preparation

The Stanford Network Analysis Project (abbreviated as SNAP) http://snap.stanford.edu/data/index.html is a huge collection of different large networks maintained by Stanford University. We downloaded a citation network dataset (High Energy Physics paper citation network cit-HepPhhttp://snap.stanford.edu/data/cit-HepPh.html), which contains 34,546 articles published in arXiv (https://arxiv.org/) during 1992-2003. The data were released in KDD Cup 2003 (http://www.cs.cornell.edu/projects/kddcup/). It has only those papers that have been cited within the network. It contains no information on whether a paper cites or is cited by another work outside the set of around 34k articles. The data and other details are publicly available on the SNAP and KDD Cup 2003 website.

We divided the citation network into subnetworks for 12 years which follows the two conditions stated below:

1. Each subnetwork for a year $n$ contains citation network for the year $n$ and $n + 1$ and
2. An edge is included between $i$ to $j$, only if $j$ is cited by paper $i$ which is published before paper $i$ in the year $n$ and $n + 1$.

The idea is that a paper's citation index is based on the recent citations numbers of that paper. The order and size of each subnetwork is shown in Table 3a.

Paper Abstracts were downloaded for all the papers (provided as KDD Supplementary Data.zip) in our dataset. Keywords were generated from the Abstracts of each paper using an efficient text mining tool named TextRank Mihalcea and Tarau (2004) in R.

### Creating Scenario Graphs

We grouped the research papers by their publication years (1992–2003) and extracted Abstracts of all the papers. Since all research papers do not have a separate keywords section, and the Abstracts have more 'terminology'-based keywords, we preferred Abstracts over keywords. On each Abstracts, a text-mining tool TextRank was used to generate a few keywords Mihalcea and Tarau (2004). Total 8538 unique keywords were obtained from all the papers through text mining. A binary matrix is created for each year with $n$ rows and $k$ columns. Each row represents a unique paper of a particular year and each column denotes a keyword. The column values over a row were taken as one, if the particular keyword is present in the paper, otherwise it is set to zero. Thus each row represents a keyword associated binary feature vector for a paper. A pairwise symmetric square distance matrix is generated using Jaccard distance from each of the binary matrices. We converted the distance matrices further into binary matrices by using a threshold value, which is the median of the all values in the respective matrix. This produced a binary matrix of the same size that was used to create a scenario graph. An edge is included in this graph (a topic similarity network in this case) if the corresponding value is one between the respective pair of nodes. The order and size of the topic similarity networks for each year are shown in Table 3(b). Apart from that, we also generated the citation subnetworks from the years 1992 to 2003.

### Analyzing the Multi-scenario Graph

We obtained year-wise topic similarity subnetworks and citation subnetworks from 1992 to 2003. For analyzing this multi-scenario graph, the GDDA was used as in the previous application. The value of GDDA score ranges within [0, 1], '1' being the highest similarity between two networks. We generated the pairwise GDDA scores between the subnetworks of both citation and topic similarity networks.

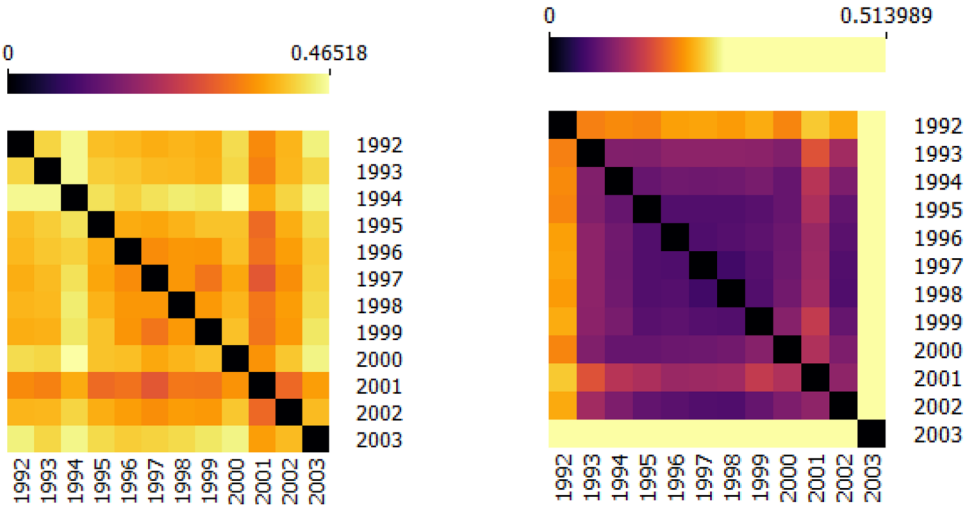



**Fig. 6** Heatmaps of the year-wise **a** citation subnetworks generated from the KDD Cup 2003 and **b** topic similarity subnetworks extracted through text mining. Both these heatmaps reflect a similar pattern

By comparing the heatmaps between the topic similarity and citation networks during 1992–2003, we observed that there is a tendency to cite the recent works among the articles. From Fig. 6a it can be seen that except for the year 1992, the trend of citing recent articles increases. This trend goes to its peak from 1995 to 1999, in the very last few years of the previous millennium. Note that, we exclude the subnetwork for the year 2003 because of having a lesser amount of data. The said network contains only 53 nodes as reported in the Table 3. The topic similarity heatmap in Fig. 6b shows more or less the same pattern over the period. There is also a topic similarity trend found during the years 1996–1999. This could be highlighting some major discoveries that happened in physics during that period.

The Fig. 7 shows the year-wise citation subnetworks. These were generated using the tool *Gephi*[1] Bastian et al. (2009). The topology that we used was the one by Yifan Hu with the following parameters: Optimal Distance—100, Relative Strength—0.2, Initial Step Size—20, Step Ratio—0.95, Convergence Threshold—0.0001.

## Conclusion and Future Works

The current paper presents an approach of analyzing complex networks constructed across multiple scenarios in a similar context. Applications of the said approach are demonstrated in two different domains. This highlights how effectively we can employ this generalized technique for any kind of scenario-based network analysis. Note that the way we portray a multi-scenario graph in the examples shown in this paper can be extended to many different applications by considering scenarios to be either location-specific or time-specific. Say, for example, if we consider that a scenario is specific to a country, we can build a scenario graph for that country depending upon the interactive activities happening within it. It could be the transportation patterns between the States (a node symbolizes a State) or may be collaboration patterns within the Universities (a node symbolizes a University). In this way, we can build a multi-scenario graph for several countries taken together for further analysis. As a limitation, we perceive that this simplistic approach can highlight only pairwise coherence between interconnected (network-based) scenarios. However, it is possible to extend this to a more complex approach with the cost of computation. Moreover, it is interesting to model this kind of approach in a streaming setting.

[1] https://gephi.org.

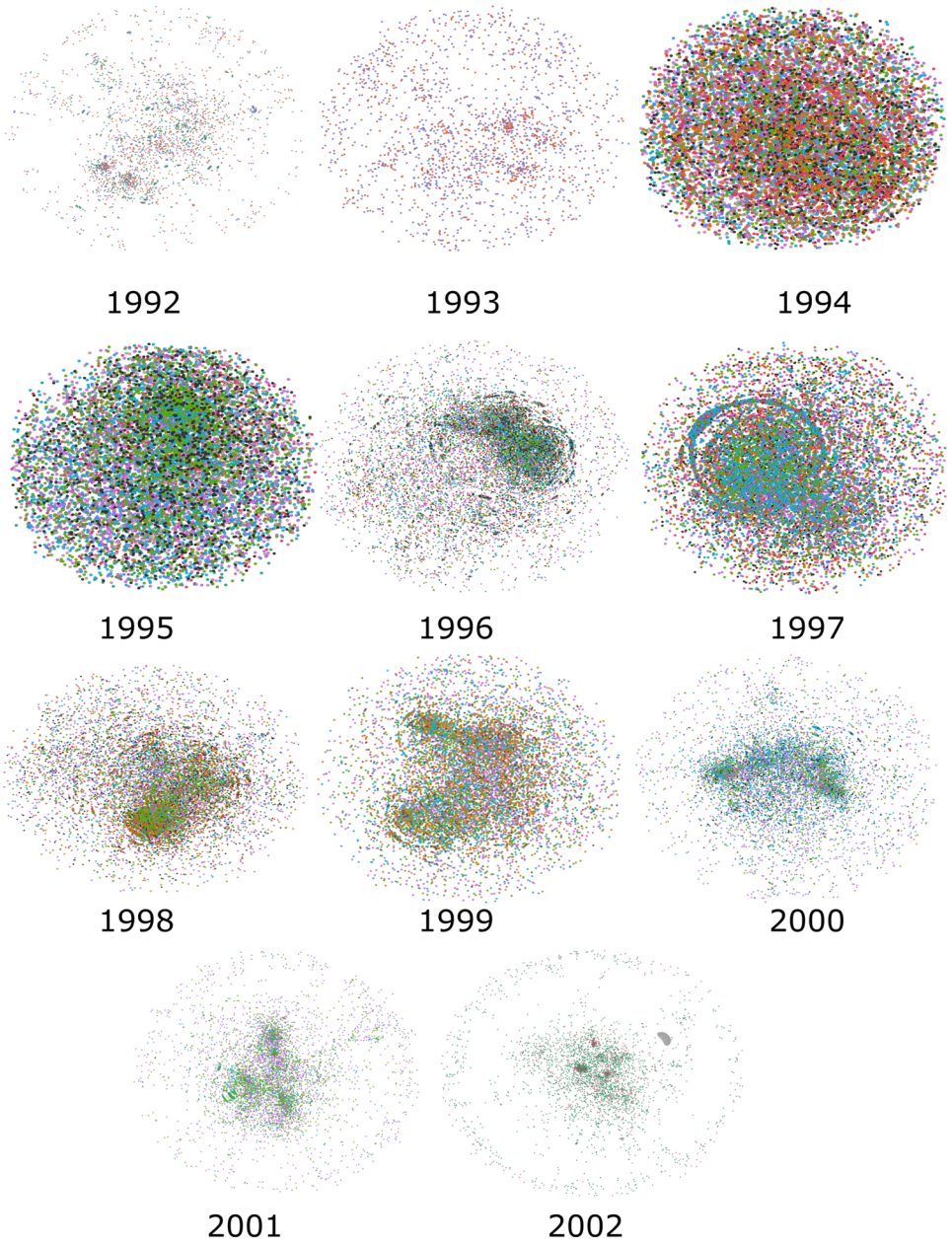



**Fig. 7** The citation subnetworks for each year during 1992–2002. A citation subnetwork specific to a year contains the papers published in that year and other papers that cited those papers. In this visualization, higher degree nodes are pushed at the center. Different colors imply the papers published in different years which cited the papers of a particular year (color figure online)

**Acknowledgements** SoB acknowledges Digital India Corporation (Formerly Media Lab Asia), Ministry Of Electronics and Information Technology (MeitY), Government of India, for providing him with a Senior Research Fellowship under the Visvesvaraya PhD Scheme for Electronics and IT. SB would like to acknowledge support from J.C. Bose Fellowship [SB/S1/JCB-033/2016] provided by the DST, Govt. of India.

**Availability** The source codes and datasets used in this paper are freely accessible from the GitHub link: https://github.com/sourav8051/Multi-scenario-Graphs-dataset.

## References

Afshinnekoo E, Bhattacharya C, Burguete-García A, Castro-Nallar E, Deng Y, Desnues C, Dias-Neto E, Elhaik E, Iraola G, Jang S et al (2021) COVID-19 drug practices risk antimicrobial resistance evolution. Lancet Microbe 2(4):e135–e136

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 3,

Biswas S, Bandyopadhyay S (2020) A Cross-Vertex Embedding Approach toward Understanding SARS-CoV-2 Variability. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)

Biswas S, Ray S, Bandyopadhyay S (2018) Analysis on Preservation Characteristics of Modular Structure during HIV-1 Progression using Weighted and Normalized Graphlet Frequency Distribution. In: 2018 4th International Conference for Convergence in Technology (I2CT), pages 1–6. IEEE

Booher DE, Innes JE (2002) Network power in collaborative planning. J Plann Educ Res 21(3):221–236

Cawkell AE (1971) Science citation index. effectiveness in locating articles in the anaesthetics field:"perturbation of ion transport''. Br J Anaesth 43(8):814

Cox MJ, Loman N, Bogaert D, O'Grady J (2020) Co-infections: potentially lethal and unexplored in covid-19. Lancet Microbe 1(1):e11

Domingo EJJH, Holland JJ (1997) RNA virus mutations and fitness for survival. Annu Rev Microbiol 51(1):151–178

Drost HG (2018) Philentropy: similarity and distance quantification between probability functions. R package version 0.2. 0. URL: https://CRAN. R-project. org/package= philentropy,

Ha S, Lee H, Choi Y, Kang H, Jeon SJ, Ryu JH, Kim HJ, Cheong JH, Lim S, Kim B-N et al (2020) Maturational delay and asymmetric information flow of brain connectivity in SHR model of ADHD revealed by topological analysis of metabolic networks. Sci Rep 10(1):1–13

Hummon NP, Dereian P (1989) Connectivity in a citation network: The development of DNA theory. Soc Netw 11(1):39–63

István F (1948) On straight-line representation of planar graphs. Acta Scientiarum Mathematicarum 11(229–233):2

Jennifer X, Chen H (2005) Criminal network analysis and visualization. Commun ACM 48(6):100–107

Kakimoto K, Kamiya H, Yamagishi T, Matsui T, Suzuki M, Wakita T (2020) Initial investigation of transmission of COVID-19 among crew members during quarantine of a cruise ship–Yokohama, Japan, February 2020

Lorts A, Smyth L, Gajarski RJ, VanderPluym CJ, Mehegan M, Villa CR, Murray JM, Niebler RA, Almond CS, Thrush P et al (2020) The creation of a pediatric health care learning network: the ACTION quality improvement collaborative. ASAIO J 66(4):441–446

Luo D, Cheng W, Yu W, Zong B, Ni J, Chen H, Zhang X (2021) Learning to drop: Robust graph neural network via topological denoising. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pages 779–787

Malay B, S Bandyopadhyay (2010) Analyzing topological properties of protein–protein interaction networks: A perspective toward systems biology. INTELLIGENCE AND PATTERN ANALYSIS IN BIOLOGICAL INFORMATICS, page 349

Mihalcea R, Tarau P (2004) TextRank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404–411, Barcelona, Spain, Association for Computational Linguistics

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC et al (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 18:1–9

Pavlopoulos GA, Wegener A-L, Schneider R (2008) A survey of visualization tools for biological network analysis. Biodata Min 1(1):1–11

Phan T (2020) Genetic diversity and evolution of SARS-CoV-2. Infect Genet Evol 81:104260

Pržulj N (2007) Biological network comparison using graphlet degree distribution. Bioinformatics 23(2):e177–e183

Shen H, Jiang Z, Zorin D, Panozzo D (2019) Progressive embedding. ACM Transactions on Graphics 38(4):32–1

Shi QL, Shun ZQ, Shen L (2019) Network Comparison Based on Structural Equivalence and Graphle, R package version 0.1.0

So MKP, Amanda AT, Chu MY, Tsang JTY, Chan JNL (2020) Visualizing COVID-19 pandemic risk through network connectedness. Int J Infect Dis 96:558–561

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z et al. (2020) On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev

Tutte WT (1963) How to draw a graph. Proc Lond Math Soc 3(1):743–767

World Health Organization et al. (2020) Naming the coronavirus disease (COVID-19) and the virus that causes it,

Yan S, Dong X, Zhang B, Zhang H-J, Yang Q, Lin S (2006) Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51

Yan S, Xu D, Zhang B, Zhang H-J (2005) Graph embedding: A general framework for dimensionality reduction. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 830–837. IEEE