Networks of Materials: Construction and Structural Analysis

Journal:	*AIChE Journal*
Manuscript ID	AIChE-20-22670
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	17-Apr-2020
Complete List of Authors:	Veremyev, Alexander; University of Central Florida, Department of Industrial Engineering and Management Systems Liyanage, Laalitha; University of North Texas, Physics Fornari, Marco; Central Michigan University, Physics Boginski, Vladimir; University of Central Florida, Department of Industrial Engineering and Management Systems Curtarolo, Stefano; Duke University Butenko, Sergiy; Texas A&M University College Station, Department of Industrial and Systems Engineering Buongiorno Nardelli, Marco; University of North Texas, Physics
Keywords:	Materials networks, similarity measures, DOS function, materials informatics, network analysis

plus1fillplus0.5fillplus0.5fillplus0.5fill

ORIGINAL ARTICLE

Journal Section

Networks of Materials: Construction and Structural Analysis

Alexander Veremyev[1] | Laalitha Liyanage[2] | Marco Fornari[3] | Vladimir Boginski[1] | Stefano Curtarolo[4] | Sergiy Butenko[5] | Marco Buongiorno Nardelli[2]

[1]Department of Industrial Engineering and Management Systems, University of Central Florida

[2]Department of Physics, University of North Texas, Denton, TX

[3]Department of Physics and Science of Advanced Materials Program, Central Michigan University

[4]Department of Mechanical Engineering and Materials Science, Duke University

[5]Department of Industrial and Systems Engineering , Texas A&M University

Correspondence
Marco Buongiorno Nardelli,
Department of Physics
University of North Texas, Denton, TX
Marco Fornari,
Department of Physics and Science of Advanced Materials Program
Central Michigan University
Email: mbn@unt.edu,
marco.fornari@cmich.edu

Funding information
U.S. Department of Defense, Office of Naval Research, Grant Number: N00014-13-1-0635

Modeling and analysis of the materials universe is an emerging area of research with many important applications in materials science. The main goal is to create a map of materials which allows not only to visualize and navigate the materials space, but also reveal complex relationships and "connections" among materials and potentially find clusters of materials with similar properties. In this paper, we consider the problem of mapping and exploring the materials universe using network science tools and concepts. The networks are based on the open-source materials data repository AFLOW.org where each material is represented as a node, and each pair of nodes is connected by a link if the respective materials exhibit a high level of similarity between their Density of States (DOS) functions. We discuss the importance of similarity measure selection, investigate basic structural properties of the resulting networks, and demonstrate advantages and limitations of the proposed approaches.

KEYWORDS
Materials networks, similarity measures, DOS function, materials informatics, network analysis, clique

1 | INTRODUCTION

Over the past several years, research in computational materials science has generated enormous amounts of data. Extracting meaningful and non-trivial insights from this data, which would reveal materials processing-structure-property-performance (PSPP) relations, is one of the main challenges in the emerging field of materials informatics, also referred to as data-driven materials science [1, 2, 3, 4, 5, 6]. In this domain it is not only important to find a material with specific properties desired in a certain application, but also to reveal relationships and "connections" between electronic structure features and to potentially identify multiple materials that have the same or similar properties of interest or are otherwise "related" according to some criteria.

The objective of this study is to contribute to research in material informatics methods by developing an approach which takes advantage of network-based modeling, analysis, and visualization techniques. Similarly to social networks, gene interaction networks and other well-known real-world complex networks, the dataset of materials can be treated as a network structure, where each individual material is represented by a node in a network, and a pair of nodes is connected by a link if two materials exhibit a certain level of similarity according to a specified quantitative measure. For example, such measures can be based on the comparison of density of states (DOS) functions, although different practical questions of interest may require the use of different similarity measures. In this approach, conceptual similarities between materials networks and social networks are clear, individual nodes are connected if they share a certain property or characteristic (i.e., materials are connected according to shared physical properties, and people are connected according to their acquaintances, collaborations, common interests, etc.).

Once a materials similarity metric is established, different approaches can be used to construct a network representation of the considered set. In an edge-weighted representation, the space of materials is treated as a complete graph (that is, all possible edges are present), and the weights of edges are given by the corresponding values of the considered similarity metrics. Alternatively, in an unweighted model, a threshold θ is selected, and only the edges of weight at least θ are assumed to be significant. Thus, in this case only pairs of materials that are considered "sufficiently similar" are linked by edges (unweighted) in the network model. This procedure is sometimes called network "slicing" [7]. Clearly, the weighted model carries more information about the materials than an unweighted one. However, there

are certain advantages of the threshold-based approach which makes it popular in network-based data mining. The main benefit is that keeping only edges indicating high level of similarity between nodes allows to apply algorithms from network analysis (specifically designed for unweighted networks) to explore its structural properties and uncover some hidden patterns and organizing principles. Another benefit is the in reduced storage requirements (which is significant given the size of the data base), which can be adjusted by selecting appropriate weight threshold values. In this paper, we focus on the threshold-based approach. Clearly, different properties of interest can determine whether a pair of nodes is connected; therefore, multiple alternative network descriptions, with different connectivity patterns, can be generated for the same set of nodes (materials).

To the best of our knowledge, the first published attempt to construct a materials network using some level of similarity between materials (fingerprints) has been presented in [8]. Specifically, the authors have introduced the notion of "material cartography" to represent AFLOW library of materials [9] as a network. They transformed the data describing band structures and density of states for each material in AFLOW into B-fingerprints and D-fingerprints, and used Tanimoto score [10] to quantify pairwise similarities among materials based on the similarity scores between the obtained fingerprints. The proposed framework was tested on more than 20,000 Inorganic Crystal Structure Database (ICSD) materials in AFLOW library and its advantages have been demonstrated for (1) searching the duplicates in AFLOW library by identifying materials with identical fingerprints and (2) identifying new compounds with interesting properties based on their similarity to known compounds (gallium arsenide, for example). However, the authors considered only similarities above 0.7 cutoff and did not report any detailed statistics of similarity score distributions or global characteristics of the corresponding materials network.

In this paper we take a further step in exploring this promising direction of research by developing a new, systematic network-based framework for mapping and structural analysis of the materials universe. Many complex systems can be analyzed via network representations, which provide a nontrivial yet intuitive mathematical tool to explore these systems. To construct network representations of materials, we examine various similarity measures commonly used in data mining applications that may capture complementary aspects of similarity between data elements. To quantify pairwise similarity between materials, we apply the selected similarity measures to density of states (DOS) functions of the respective materials (complete or partial DOS function data can be used). In particular, we compute similarity

measures for all pairs of materials in the dataset of around 27,000 ICSD materials with their DOS functions computed and stored in the AFLOW data repository. We analyze the distributions of the obtained similarity metrics, discuss some advantages and disadvantages of the considered measures, and develop new, enhanced metrics, based on weighted Pearson correlation coefficient with some extra adjustments. Furthermore, we show that the constructed networks exhibit pronounced "small-world" properties, which are strikingly similar to many other real-world networks.

As it was mentioned in [8], the similarity concept can be used as an effective tool for searching materials with similar properties in large databases. However, in addition to exploring pairwise similarities of materials, one may be interested in identifying large groups of materials sharing similarity according to some property. If the objective is to have a high similarity score for any pair of materials within the group, this group would correspond to a *clique* in the network constructed based on the considered similarity criterion. A clique is defined as a subset of nodes that are all adjacent to each other [11]. The clique concept is used in numerous application areas due to its elegance and inherent ability to logically represent cohesive subgroups of "tightly knit" elements (i.e., nodes) in complex systems modeled as graphs [12]. For example, in social networks, where the vertices correspond to "actors" and an edge indicates a relationship between two actors [13], a clique represents a group of people any two of which have a certain kind of relationship (friendship, acquaintance, etc.) with each other [14]. In fact, some of the earliest work on cliques and methods of their detection was motivated by applications in sociometry [11, 15, 16]. Hence, in this paper we are also interested in computing cliques in the constructed materials networks. In particular, we will find maximum cliques in materials networks based on various similarity cutoffs, which provide a "global" cohesiveness characterization for the whole network. In addition, we will compute the largest cliques containing certain materials selected for our "local" analysis. Finally, we demonstrate a prototype network-based navigation tool we developed for the AFLOW library.

The remainder of this paper is organized as follows. Sec. 2 describes the proposed methodology for constructing materials networks according to DOS-based similarity metrics. Sec. 3 presents the obtained results, including the similarity score distributions, global and local characteristics of the constructed materials networks, as well as prototype network-based navigation tool for AFLOW library. Sec. 4 concludes the paper with a discussion of future research directions.

2 | METHODOLOGY

2.1 | Dataset and Preprocessing

The experiments are performed on the dataset containing DOS values of 27,007 ICSD materials obtained from AFLOW library available at AFLOW.org [9]. AFLOW is one of the largest electronic structure data repositories (>1,100,000 entries) in the world, designed specifically with materials informatics in mind.

Each material in our dataset is represented by its DOS function consisting of a list of 668 values. The first half (334 values) corresponds to the DOS function values calculated at evenly distributed points (0.015eV apart) over the energy from -5.0 eV to E_v, where E_v is the top of the valence band of the Fermi energy. The remaining half (334 values) corresponds to the DOS function values above the energy for E_c (the bottom of the conduction band) or the Fermi energy over to 5 eV range. DOS functions of both spin polarized and non spin polarized calculations are considered. In the case of spin polarized calculations spin up and spin down DOS are added together.

The DOS function values over energy regions below and above Fermi energy are referred to as negative and positive energy regions, respectively. This representation was chosen (1) to deal with functional properties influenced by features of the valence band or of the conduction band separately, and (2) to avoid uncertainties associated with the calculated energy gap ($E_c - E_v$) that is recorded but not used in the metric. In the following we compute similarity scores between their DOS functions over negative and positive energy regions. The values of DOS functions are scaled in such a way that their sums over negative and positive energy regions are equal to 100. This choice biases toward the presence of feature in the DOS over the magnitude of such features, but allows to distill some of the main characteristics of the band structure of a material into a single descriptor and build meaningful relations, to learn regarding a variety of physical properties, such as transport, optical response, etc.

As an example, the material Si (ICSD 150530) has the band structure and the DOS function illustrated on Fig. 1 (obtained from AFLOW library [1]). The corresponding vectors of scaled DOS values over positive and negative energy regions considered in this paper for similarity computations are depicted on Fig. 2. pdf Note that this material has a theoretical band gap 0.675 eV. Hence, the i-th component value of DOS vector over negative energy region corresponds

[1] http://aflowlib.org/material.php?id=150530

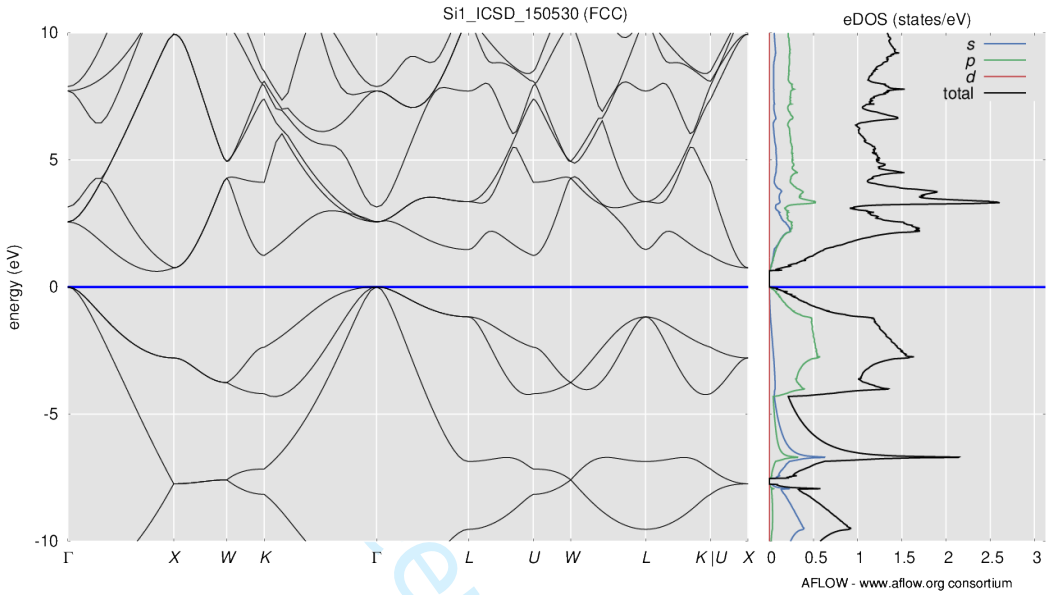
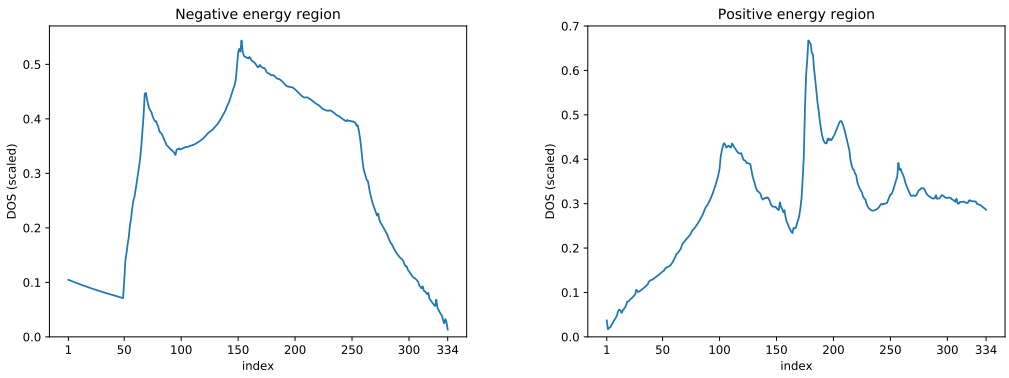


FIGURE 1 Band structure and DOS of Si (ICSD 150530) obtained from AFLOW library.



(a) (b)

FIGURE 2 Vectors of DOS values over (a) negative [-5,0] and (b) positive [0,5] energy regions.

to the DOS function value at $-0.015 \times (334 - i)$ eV and the i-th component of DOS vector over positive energy region corresponds to the DOS function value at $0.675 + 0.015 \times (i - 1)$ eV.

In total, for a given similarity metric, the corresponding similarity score has been calculated for 364,657,521 pairs of materials. Moreover, our dataset contains information about the band gap for each material (since the band gap is not used in the similarity score computation), which we use to separate materials into three classes (metals, semiconductors, insulators).

2.2 | Similarity Measures

In this section, we first review the most common (symmetric) similarity measures used in various data-mining applications (primarily for comparing probability distributions as DOS is essentially a distribution function) to quantify the proximity/similarity among data points represented by vectors. We discuss their advantages and disadvantages in terms of capturing certain aspects of physical similarity among materials encoded in their DOS functions. Intuitively, the more similar the DOS values of two materials, the more similar properties the respective materials should have. However, as it will be discussed later, the shape of DOS function and its behavior in a certain energy region (around Fermi level) plays a key role in determining physical properties of the corresponding material. Therefore, for the network construction and analysis, we develop a similarity measure which takes this consideration into account and is able to better reflect the similarity among materials properties.

For a pair of materials, let vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be their DOS function values, respectively, over the positive or negative energy region ($n = 334$ in our experiments). We consider six similarity metrics: two intersection-based (Jaccard similarity and average ratio), two distance-based (Euclidean distance and Manhattan distance), and two inner product-based (cosine similarity and Pearson correlation coefficient). These similarity metrics are chosen due to their frequent use in data-mining applications (see, e.g., [17] for a survey of various similarity metrics among probability distributions and their practical usage). A brief description of each of the considered measures is given next. Note that four of the considered similarity measures (Jaccard, average ratio, cosine, and Pearson correlation coefficient) are "score"-based, that is, the larger the computed value of the respective similarity measure between a pair materials, the more similar these materials are. On the contrary, the remaining two similarity measures (Euclidean and Manhattan distances) are "distance"-based, with smaller values of the similarity measure implying that materials

are "close" to each other in the materials space or have high levels of similarity. To distinguish between these types of similarity measures, score-based measures are denoted by S, whereas distance-based measures are denoted by D.

Jaccard similarity

This is essentially a measure of relative *overlap* of areas under the DOS functions of materials represented by vectors x and y. The Jaccard similarity score $S_J(x, y)$ is given by

$$S_J(x, y) = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)}. \quad (1)$$

In [17] it is also referred to as Ruzicka similarity, or one minus Tanimoto (a.k.a Jaccard) or Soergel distance. It has been used in [8] to construct and map networks of materials based on similarity among their fingerprints.

Average ratio

This measure provides the average ratio between minimum and maximum values of DOS functions at n considered energy levels:

$$S_a(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{\min(x_i, y_i)}{\max(x_i, y_i)}. \quad (2)$$

It corresponds to one minus Wave Hedges distance, mentioned in [17].

Euclidean distance (L_2)

This measure treats x and y as points in n-dimensional space R^n and computes a (scaled) L_2 distance between these two points:

$$D_E(x,y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{3}$$

We scale it because one can use DOS functions represented by various number of points and we would like to ensure that this measure is not sensitive to the density of the selected grid (points per unit of energy).

Manhattan distance (L_1)

This metric also treats x and y as points in n-dimensional space R^n and computes a (scaled) L_1 distance between these two points, which is less sensitive to outliers (points with larger difference in DOS function values):

$$D_M(x,y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|. \tag{4}$$

Cosine similarity

This measure considers x and y as vectors in n-dimensional space R^n and computes a cosine value of an angle between these two vectors:

$$S_c(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{5}$$

Cosine similarity is a popular vector based similarity measure in text analytics and information retrieval applications (see e.g., [18]).

Pearson correlation coefficient

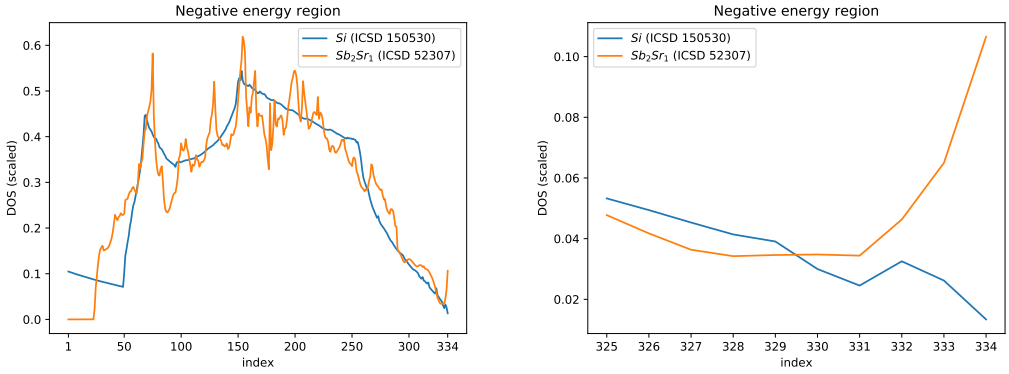
This measure quantifies the linear correlation between two vectors x and y:

$$S_p(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \quad (6)$$

Pearson correlation coefficient is essentially a cosine value of an angle between centered versions of these two vectors x and y, i.e., vectors with subtracted mean values. Pearson correlation coefficient along with cosine similarity are most widely used and popular similarity measures.

The results of preliminary experiments we conducted have shown that none of the six standard similarity measures described above yielded satisfactory results from a physical perspective. More specifically, we observed that for some pairs of materials that significantly differed in terms of physical properties of interest, their similarity score was unexpectedly high. Only Pearson correlation coefficient seemed to have promising results and was better in capturing the similarity among DOS tipping points and surrounding areas which are important in determining materials properties.

The main reason these similarity measures among DOS function values might not be able to adequately capture the similarity of materials properties is that the materials properties are primarily encoded not only in the DOS function values over all possible energy values, but also in the shape and the behavior of respective DOS function values around zero (Fermi) energy values (or right after the band gap region). Specifically, the density of states values and the rate at which these values change are a direct consequence of the details of the band structure of the material, and thus the important factors determining materials properties, such as electronic transport and optical response. Moreover, for all practical applications, the farther the considered energy level from Fermi energy, the less important are the density of states values in defining the materials properties.



(a) (b)

FIGURE 3 Vectors of DOS values of Si (ICSD 150530) and Sb2Sr1 (ICSD 52307) over (a) 334 points and (b) 10 points near zero in negative energy region.

For example, according to Jaccard similarity measure over negative energy region, one of the most similar materials to Si (ICSD 150530) is Sb2Sr1 (ICSD 52307) with similarity score 0.85, which essentially means that the DOS functions have roughly 85% overlap (Fig. 3a). However, one can see that DOS close to zero energy have different behavior (Fig. 3b). Not surprisingly, the material Sb2Sr1 (ICSD 52307) is substantially different from Si (ICSD 150530); it appears to be a metal and has a more complex band structure.

Taking these considerations into account, we develop a similarity measure which is based on Pearson correlation coefficient and introduce some adjustments to make it more suitable for capturing materials similarity.

Weighted Pearson correlation coefficient

Since density of states values become less important (for determining materials properties) as they get farther away from the Fermi energy, we introduce *weight functions* that put higher weights on the values close to the Fermi energy. Based on our preliminary numerical experiments, the following weight functions, depicted on Fig. 4 appear to be reasonable to address this concern. Specifically, the considered energy region is divided into five approximately equal point sets, and the weight w_i is set to 1 for the points near zero energy (for negative energy region) or right after the band gap (for positive energy region), and then it drops by 0.2 each time we move farther from that set. Then the

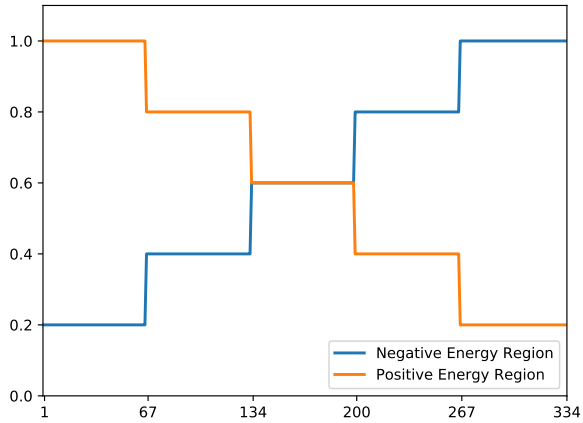


FIGURE 4 Weight functions over positive and negative energy regions for Pearson coefficient computations.

weighted Pearson correlation coefficient for vectors x and y is expressed by the following formula:

$$S_{wp}(x,y) = \frac{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sqrt{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2}\sqrt{\sum_{i=1}^{n} w_i(y_i - \bar{y}_w)^2}}, \tag{7}$$

where $\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ and $\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$.

Furthermore, to capture the similarity of DOS behavior near zero energy (or right after the band gap region), we introduce two adjustments. Specifically, they are based on the last (for negative energy region) or the first (for positive energy region) k components of the DOS vector where k is a small integer.

Adjustment 1

The first one measures the similarity between the averages of DOS function values over these k components. It is equal to the ratio of the lowest average to the highest one among these two. It is similar in spirit to two ratio-based similarity

measures mentioned above, i.e., Jaccard similarity or average ratio. The explicit equations for this adjustment over negative and positive energy region are given next.

For the negative energy region we have the following formula:

$$A_{neg}(x,y,k) = \frac{\min\left(\sum_{i=n-k+1}^{n} x_i, \sum_{i=n-k+1}^{n} y_i\right)}{\max\left(\sum_{i=n-k+1}^{n} x_i, \sum_{i=n-k+1}^{n} y_i\right)}. \tag{8}$$

For the positive energy region we obtain:

$$A_{pos}(x,y,k) = \frac{\min\left(\sum_{i=1}^{k} x_i, \sum_{i=1}^{k} y_i\right)}{\max\left(\sum_{i=1}^{k} x_i, \sum_{i=1}^{k} y_i\right)}. \tag{9}$$

Adjustment 2

The second adjustment measures the similarity between the changes of respective DOS functions. Specifically, for a pair of DOS functions represented by vectors x and y, it is based on differences among angles of DOS value changes from point 1 to point k. It is computed using the following equations.

For the negative energy region we have:

$$\alpha_{neg} = \arctan\left(\frac{x_{n-k+1} - x_n}{(k-1)e_0}\right), \tag{10}$$

$$\beta_{neg} = \arctan\left(\frac{y_{n-k+1} - y_n}{(k-1)e_0}\right), \tag{11}$$

$$B_{neg}(x,y,k) = 1 - \frac{|\alpha_{neg} - \beta_{neg}|}{\pi/2}, \tag{12}$$

where $e_0 = 5/(n-1)$ is the energy range between two consecutive data points (i.e., 0.015eV in our experiments).

Similarly, for the positive energy region, we have:

$$\alpha_{pos} = \arctan\left(\frac{x_k - x_1}{(k-1)e_0}\right), \tag{13}$$

$$\beta_{pos} = \arctan\left(\frac{y_k - y_1}{(k-1)e_0}\right), \tag{14}$$

$$B_{pos}(x, y, k) = 1 - \frac{|\alpha_{pos} - \beta_{pos}|}{\pi/2}. \tag{15}$$

Weighted Pearson correlation coefficient with (0.5,0.5) adjustments

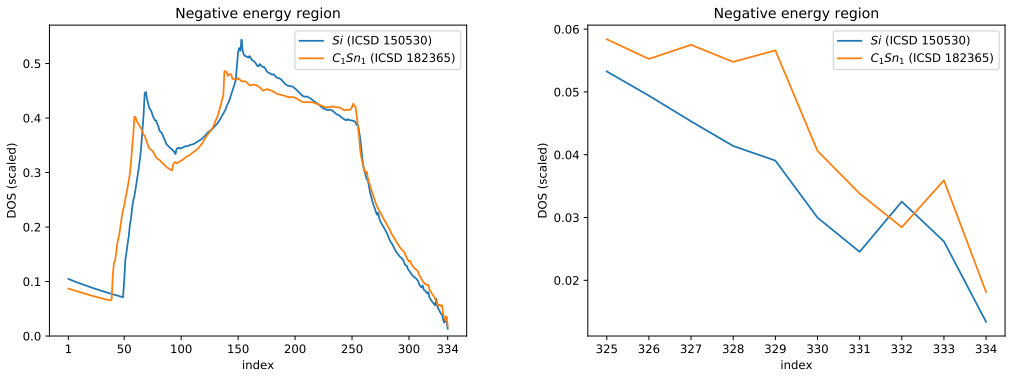
Our final proposed similarity measure is comprised of weighted Pearson correlation coefficient multiplied by the weighted sum of the two adjustments. Our findings indicate that equally weighted adjustments provide satisfactory results:

$$S_{wp}^{adj}(x, y) = S_{wp}(x, y) \times (0.5A + 0.5B), \tag{16}$$

where $A = A_{neg}(x, y, k)$ and $B = B_{neg}(x, y, k)$ for the negative energy region, and $A = A_{pos}(x, y, k)$ and $B = B_{pos}(x, y, k)$ for a positive energy region. Based on our preliminary computational experiments, we set $k = 3$ for Adjustment 1 and $k = 5$ for Adjustment 2. Note that if $x = y$, then $S_{wp}^{adj}(x, y) = 1$, hence close to 1 similarity scores should indicate high level of similarity. We use this measure for the construction and analysis of materials network in the experiments reported in this paper.

To illustrate that this measure provides reasonable results, consider the most similar material to Si (ICSD 150530) according to this measure, which is C1Sn1 (ICSD 182365) with 0.89 similarity score S_{wp}^{adj} over negative energy region (Table 3, discussed in more detail below). It is based on 0.982 weighted Pearson correlation coefficient, 0.874 adjustment 1 and 0.944 adjustment 2 scores. One can clearly see that not only overall behaviors of DOS functions are similar (Fig.

5a), but also DOS close to zero energy have similar behaviors as well (Fig. 5b). These two materials can indeed be considered as similar since they belong to same group in the periodic table and have the same crystal structure.



(a) (b)

FIGURE 5 Vectors of DOS values of Si (ICSD 150530) and C1Sn1 (ICSD 52307) over (a) 334 points and (b) 10 points near zero in negative energy region.

We also note that the material Sb2Sr1 (ICSD 52307) we mentioned before (to illustrate the drawbacks of using Jaccard similarity score), has only 0.28 similarity score according to S_{wp}^{adj}, which is primarily due to the difference in the DOS behavior near zero energy (adjustment 1 is equal to 0.331 and adjustment 2 is equal to 0.271).

2.3 | Constructing Networks of Materials

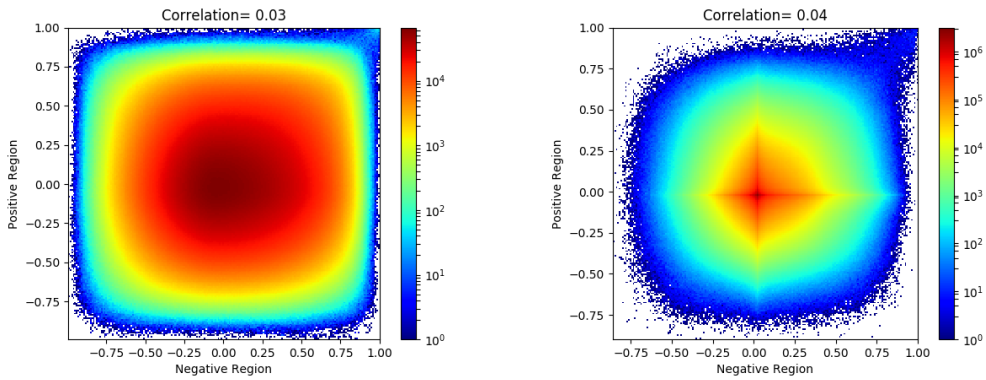
Similarly to social, biological, technological and other well-known real-world complex networks, the dataset of materials can be treated as a network, where each individual material is represented by a node, and a pair of nodes is connected by an edge (link) if the respective two materials exhibit a certain level of similarity according to a specified quantitative measure. Using a similarity measure, such as those mentioned above, one can construct a materials network $G = (V, E)$, where V denotes the set of nodes and $E \subseteq \binom{V}{2}$ is the set of edges, as follows. For each pair of materials x and y in V, given a similarity measure $S(x, y)$ (or $D(x, y)$) and a threshold value C, we create a link $\{x, y\} \in E$ between nodes x and y if $S(x, y) \geq C$ (or $D(x, y) \leq C$). Clearly, each distinct choice of a similarity measure and the corresponding threshold

value would produce a distinct instance of a network of materials. In the next section, we analyze the constructed network instances and present the obtained results.

3 | RESULTS

In this section, we present the results of experiments concerning network representations of the AFLOW library data using the introduced weighted Pearson correlation similarity measure (7) with adjustment (16).

3.1 | Similarity Score Distributions



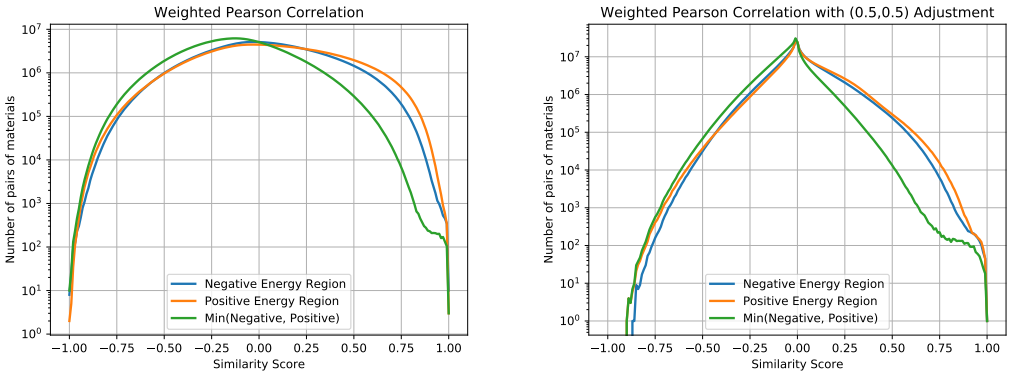
(a) (b)

FIGURE 6 Heat maps of similarity scores computed over positive and negative energy regions based on (a) weighted Pearson correlation coefficient S_{wp} and (b) weighted Pearson correlation coefficient with adjustment S_{wp}^{adj}

First, we provide the plots illustrating similarity score distributions computed for all pairs of materials. Specifically, since for each pair of materials the corresponding similarity scores over negative and positive energy regions are computed separately, Fig. 6 shows the heat maps of similarity scores (weighted Pearson correlation coefficient without and with adjustment) between DOS functions over negative (horizontal axis) and positive (vertical axis) energy regions, as well as correlation coefficients for all pairs of materials in our dataset. Our findings indicate that there is no clear

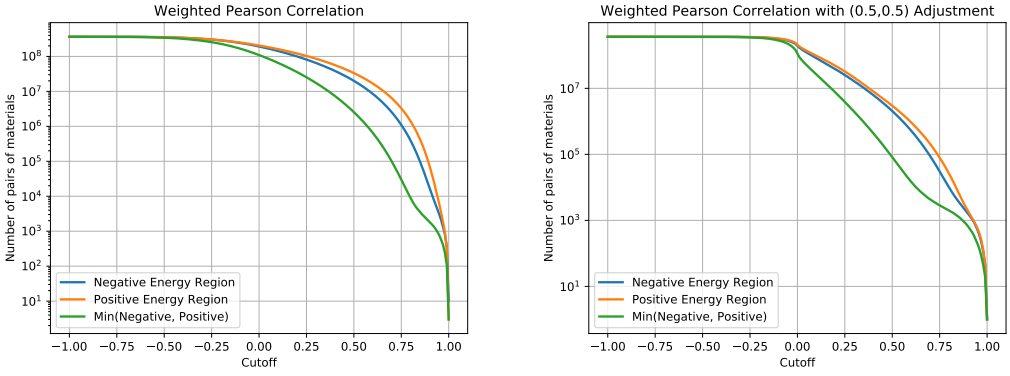
correlation between similarity scores over positive and negative energy regions, which means that the similarity score of two materials over one region does not correlate with the similarity score over the other region. It also supports our motivation for considering similarity over positive and negative energy regions separately as the behavior of DOS over these two regions reflects different materials properties. However, as one can observe on the top right corner of Fig. 6b, if the similarity score over negative energy region is close to 1, it is more likely that the similarity score over positive energy region is also close to 1, i.e., if two materials DOS are very similar in one energy region, it is more likely that their DOS are very similar over the other energy region as well.

Figures 7 and 8 illustrate the distributions of similarity scores (weighted Pearson correlation coefficient without and with adjustment) over positive and negative energy regions. In addition, we plot the distribution of minimum among the scores over negative and positive energy regions. One can observe a slight variation within the distribution in the sense that the number of pairs of materials with high similarity scores (e.g., > 0.5) over positive energy region is greater than that over negative energy region. The peak is observed near the zero similarity score, which means that the majority of pairs of materials do not have similar DOS functions. From the networks perspective, our cutoff region of interest is where the number of pairs of materials is not very large (somewhere around 10^6, which corresponds to cutoff above 0.6), such that the constructed networks are not very dense, meaningful, and reflect some structural



(a) (b)

FIGURE 7 Number of materials with given similarity scores: (a) weighted Pearson correlation coefficient S_{wp} and (b) weighted Pearson correlation coefficient with adjustment S_{wp}^{adj}. The number of materials is taken over 0.01 region.



(a) (b)

FIGURE 8 Cumulative distribution of similarity scores: (a) weighted Pearson correlation coefficient S_{wp} and (b) weighted Pearson correlation coefficient with adjustment S_{wp}^{adj}.

aspects of the corresponding materials space. Note that the authors of [8] use 0.7 cutoff for network construction and visualization, although they do not provide any details on the choice of such threshold, and the similarity score they are using is somewhat different.

3.2 | "Global" Characteristics of Networks of Materials

In this set of experiments we construct materials networks using various threshold levels and investigate their basic structural properties. Fig. 9 illustrates the distributions of the degrees in materials networks in their largest connected components for four threshold levels (0.8, 0.75, 0.7, 0.65) constructed using the adjusted weighted Pearson correlation coefficient (16) as the similarity metric. Specifically, each marker on a plot represents the number of materials (horizontal axis) with a certain degree (vertical axis) in log-log scale. We would like to emphasize that similarly to many real-life networks, the degree distribution of materials networks (for any threshold) resembles a straight line, which indicates that the degree distributions are somewhat similar to those described by a power law. Note that the degree distribution for the largest connected component has *low-degree saturation*, meaning that low-degree nodes are less frequent than what is predicted by the power law. There is also a high-degree cutoff, indicating that there are fewer high-degree nodes

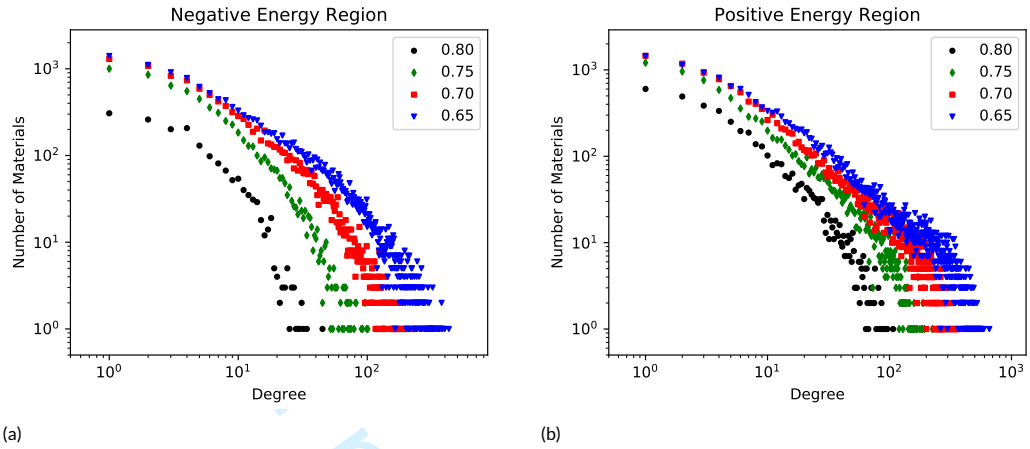


FIGURE 9 Distributions of the degrees in the largest connected components of materials networks constructed using adjusted weighted Pearson coefficient S_{wp}^{adj} over (a) Negative energy region and (b) Positive energy region

than what is expected in a pure power law. These observations are common deviations from power law behavior in real networks [19].

Tables 1 and 2 report the basic statistics of the materials networks constructed over the negative and positive

TABLE 1 Basic characteristics of the materials networks constructed based on the weighted Pearson coefficient with adjustments over the negative energy region.

Threshold	$\|E\|$	$\|V_1\|$	$\%\|V_1^m\|$	$\%\|V_1^s\|$	$\%\|V_1^{ins}\|$	$\|V_2\|$	$\#isol$	$\#comp$	$\overline{Deg}$	$Diam$	$AvgDist$	C_1	C_2	$DegCorr$	$MaxCl$
0.9	1578	16	100	0	0	15	25219	662	3.875	5	2.275	0.622	0.411	0.254	6
0.85	3690	71	98.6	0	1.4	48	23523	1103	4.648	15	5.49	0.639	0.396	0.646	10
0.84	4394	131	99.2	0.8	0	120	23116	1149	3.389	26	10.146	0.47	0.356	0.181	7
0.83	5279	306	98.4	1.6	0	299	22620	1192	4.275	24	8.218	0.506	0.413	0.324	11
0.82	6467	610	98.7	1.1	0.2	480	22071	1197	4.659	27	9.811	0.462	0.38	0.367	11
0.81	7931	1130	92.4	6.4	1.2	561	21507	1221	5.166	39	12.546	0.484	0.414	0.379	13
0.8	9871	1692	91.9	6.1	2	713	20928	1214	5.457	41	11.97	0.436	0.389	0.373	13
0.79	12357	3061	90	5	5	90	20284	1180	5.823	56	17.473	0.427	0.401	0.409	15
0.78	15363	3755	88.3	6.7	5	164	19661	1120	6.351	63	16.175	0.411	0.394	0.423	15
0.77	19248	4783	83.2	8.4	8.4	64	18973	1095	6.799	55	14.79	0.396	0.392	0.458	16
0.76	24268	5593	81	9.7	9.3	78	18278	1053	7.638	53	12.961	0.383	0.392	0.466	18
0.75	30381	6467	79.6	10.5	9.9	32	17592	1017	8.584	47	12.147	0.376	0.401	0.475	19
0.74	37921	7306	77.6	11	11.3	36	16871	992	9.694	42	11.415	0.37	0.401	0.476	21
0.73	47268	8193	76.7	11.3	12.1	32	16144	960	10.962	44	10.972	0.366	0.402	0.483	23
0.72	58467	9016	75.2	11.7	13.1	33	15409	920	12.464	39	10.201	0.365	0.401	0.485	26
0.71	72144	9959	73.9	12.2	13.9	29	14674	872	14.074	43	9.721	0.363	0.4	0.482	28
0.7	88495	10797	72.7	12.5	14.7	16	13948	836	16.04	34	9.105	0.363	0.411	0.479	31
0.69	107778	11661	71.8	12.7	15.5	14	13245	789	18.185	36	8.673	0.363	0.413	0.48	33
0.68	130960	12485	70.7	13	16.3	13	12541	749	20.719	34	8.194	0.364	0.421	0.479	39
0.67	158403	13245	69.7	13.2	17.1	20	11890	702	23.688	31	7.807	0.364	0.426	0.476	44
0.66	190692	14084	68.8	13.4	17.7	20	11247	635	26.885	30	7.542	0.366	0.427	0.473	49
0.65	228035	14853	67.9	13.7	18.3	20	10603	602	30.539	29	7.267	0.367	0.432	0.469	52

TABLE 2 Basic characteristics of the materials networks constructed based on the weighted Pearson coefficient with adjustments over the positive energy region.

Threshold	$\|E\|$	$\|V_1\|$	$\%\|V_1^m\|$	$\%\|V_1^s\|$	$\%\|V_1^{ins}\|$	$\|V_2\|$	$\#isol$	$\#comp$	$\overline{Deg}$	$Diam$	$AvgDist$	C_1	C_2	$DegCorr$	$MaxCl$
0.9	1871	46	0	19.6	80.4	45	25026	644	2.565	10	4.196	0.195	0.203	-0.235	3
0.85	6810	704	1.1	29.4	69.5	458	22564	1034	6.33	29	8.462	0.357	0.363	0.223	9
0.84	8994	1337	2.8	32	65.2	600	21893	1079	6.402	40	11.465	0.377	0.342	0.315	11
0.83	11880	1614	3	33.3	63.6	754	21218	1119	7.502	36	10.497	0.385	0.372	0.344	12
0.82	15469	1869	3.4	34.1	62.5	1097	20542	1129	8.748	33	9.666	0.393	0.383	0.364	13
0.81	20070	2117	3.4	34.8	61.8	1418	19859	1143	10.334	31	9.079	0.401	0.404	0.375	14
0.8	25642	4045	43.6	21.3	35.1	216	19137	1163	10.536	60	17.436	0.401	0.411	0.417	18
0.79	32688	4833	47.3	19.8	32.8	270	18456	1139	11.935	61	15.984	0.404	0.419	0.423	21
0.78	41381	5684	51.2	18.2	30.7	300	17670	1141	13.213	73	16.858	0.406	0.42	0.43	23
0.77	52014	6415	53	17.6	29.4	518	16875	1119	14.954	54	14.814	0.407	0.426	0.43	26
0.76	64261	7762	53.1	17.2	29.7	34	16074	1112	15.861	56	16.139	0.408	0.421	0.445	30
0.75	79503	8766	54	17	29	20	15257	1066	17.559	58	15.154	0.409	0.416	0.443	32
0.74	97698	9715	54.5	16.7	28.8	32	14523	994	19.629	53	14.136	0.411	0.423	0.441	35
0.73	118658	10659	55.4	16.5	28.1	25	13676	965	21.849	48	13.169	0.414	0.424	0.442	42
0.72	143579	11652	55.6	16.4	28	16	12971	891	24.323	54	12.628	0.416	0.43	0.44	46
0.71	172937	12517	56.1	16.6	27.4	12	12245	837	27.349	43	11.773	0.418	0.43	0.439	53
0.7	206139	13333	56.2	16.4	27.4	19	11519	793	30.664	38	10.995	0.42	0.432	0.44	58
0.69	244849	14198	56.3	16.3	27.4	12	10812	735	34.264	34	10.359	0.423	0.431	0.438	61
0.68	289260	14976	56.1	16.2	27.7	12	10130	696	38.423	32	9.814	0.425	0.436	0.436	67
0.67	339650	15777	55.9	16.2	27.9	12	9516	636	42.879	30	9.281	0.427	0.442	0.433	75
0.66	397184	16516	55.8	16.2	28	12	8802	627	47.93	28	8.727	0.43	0.445	0.432	83
0.65	462140	17263	55.9	16	28	12	8204	568	53.396	28	8.342	0.433	0.446	0.431	88

energy regions, respectively. Each of the networks has $|V|$=27,007 nodes. Out of 27,007 corresponding materials, 14,262 (53%) are metals, 4,156 (15%) are semiconductors, and 8,588 (32%) are insulators. We treat a material as a metal if it has 0 band gap, as a semiconductor if it has a positive band gap less than 1.5 eV, and as an insulator, if its band gap is above 1.5 eV. The notations used to describe the columns in these tables are as follows: $|E|$ – the number of edges; $|V_1|$ – the number of nodes in the largest connected component; $\%|V_1^m|, \%|V_1^s|, \%|V_1^{ins}|$) – the percentages of 'metals', 'semiconductors' and 'insulators' in the largest connected component, respectively; $|V_2|$ – the number of nodes in the second largest connected component; $\#isol$ – the number of isolated nodes; and $\#comp$ – the number of connected components (disregarding the isolated nodes). The remaining columns contain characteristics of the largest connected component: $\overline{Deg}$ – the average degree; $Diam$ – the diameter; $AvgDist$ – the average distance; C_1 – the global clustering coefficient; C_2 – the average local clustering coefficient; $DegCorr$ – the degree correlation; and $MaxClique$ – the size of the maximum clique.

Based on these tables, it is interesting to observe that materials networks appear to be small world networks [20], which means that they have high clustering coefficients, small diameters and small average distances. Moreover, average distance, clustering coefficients and degree correlation of the materials network constructed with, e.g., 0.7 cutoff, are

surprisingly similar to the corresponding parameters of collaboration network among physicists (two physicists are connected by an edge if they co-authored at least one paper) presented in Table II in [21]. Namely, the materials network has average distance 6, clustering coefficients 0.37 and 0.51, and degree correlation 0.27, whereas the same parameters for the physicists' collaboration network are equal 6.19, 0.45, 0.56 and 0.363, respectively. Note that other networks studied in the aforementioned paper have somewhat different values of such parameters.

Another interesting observation is that the second largest connected component is usually very small in comparison to the largest one. It means that as the threshold level decreases from 0.8 to 0.65, only one giant component emerges (containing mostly metals), and the rest of the network is mostly comprised of isolated nodes and very few connected components of small sizes. However, as the largest connected component grows, it contains more and more semiconductors and insulators. For example, for a threshold level 0.65, out of 27,007 materials there is one connected component with 14,673 nodes (70.9% metals, 14.4% semiconductors and 14.6% insulators), 10,214 isolated nodes, and 827 connected component of small sizes, with the total of 2,120 nodes (average size < 3).

This observation about uneven distribution of metals, semiconductors and insulators in the connected components indicates that different types of materials (metals, semiconductors or insulators) should be treated differently when choosing a similarity threshold for any particular application. The results presented in the next section will further support this conclusion.

Note that in Tables 1 and 2 we also report the results on cardinality of maximum cliques in the largest connected components of the corresponding materials network. For example, the largest clique of size 10 for the 0.85 cutoff (Table 1) contains the following set of materials: Na1 (ICSD 426957), Li1 (ICSD 642102), Ca10·6758 (ICSD 163535), S1Tl1 (ICSD 52201), Li1 (ICSD 642105), Ba1Li4 (ICSD 615944), Ho1 (ICSD 639322), O1Ta1 (ICSD 647483), Ca10·7184 (ICSD 163531), and Na1 (ICSD 53753). All materials in this set are metals with occupied states bands that are largely parabolic.

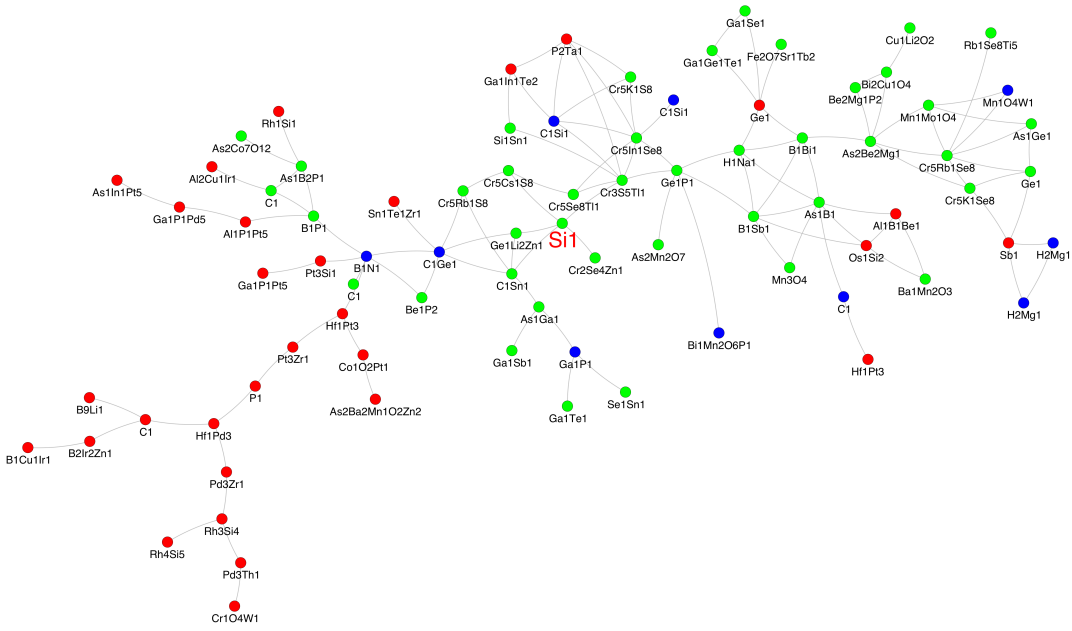


FIGURE 10 Largest connected component of materials network that includes Si (ICSD 150530) containing 82 materials and 116 edges obtained for 0.82 cutoff of weighted Pearson coefficient with adjustment computed over negative energy region. The nodes are colored according to the band gap: red - metals (0 eV), green - semiconductors (0-1.5 eV), blue - insulators (>1.5 eV).

3.3 | "Local" Characteristics of Networks of Materials

While global structural properties of a materials network may reveal some basic collective characteristics of the materials universe emerging from the information about pairwise similarities of the materials, in practice one is often interested in a "local neighborhood" of a particular material, that is, a group of materials that are similar to the given material. In this section we present sample results on local analysis for the constructed materials networks. Specifically, to illustrate the approach, we analyze the local neighborhoods of the following 4 prototypical materials: Si (ICSD 150530), the most common semiconductor; Al (ICSD 43492), a typical metal; Ba1O3Ti1 (ICSD 183932), a wide band-gap insulator of technological interest; and Co1Sb3 (ICSD 164980), a well known thermoelectric of a particularly complex band structure. Fig. 10 illustrates the largest connected component of the materials network that includes Si (ICSD 150530) constructed for the adjusted weighted Pearson coefficient similarity score between DOS function values over all energy regions combined, with 0.82 cutoff.

TABLE 3 Top 20 similar materials according to our similarity score to Si (ICSD 150530) and Al (ICSD 43492) over negative and positive energy regions.

	Si (ICSD 150530)						Al (ICSD 43492)					
	Negative			Positive			Negative			Positive		
Rank	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score
1	C1Sn1	182365	0.893	Ag1Mg1Sb1	187151	0.861	Dy1Zn5	630386	0.728	Ga1	12174	0.919
2	Cr5Cs1S8	2566	0.842	Cu3N1	53313	0.777	C1Ta1	185986	0.705	P1	98121	0.839
3	Ge1Li2Zn1	171498	0.825	Ag1Bi1Ca1	659377	0.742	Ga1	12174	0.691	Hg1Mg1	639081	0.838
4	Cr2Se4Zn1	626758	0.825	Sb4Sn3Sr1	165617	0.73	Y1Zn5	106228	0.69	Si1	109025	0.814
5	Cr3S5Tl1	23632	0.825	As2Ba2Mn1O2Zn2	85659	0.709	Cd1Ge1P2	52804	0.673	Si1	41392	0.811
6	Cr5Se8Tl1	37123	0.812	As2Nb1	18143	0.695	La1Zn5	104736	0.671	Ni1Si2Tb1	54298	0.81
7	Cr5Rb1S8	2567	0.802	As4Ba3Zn2	424760	0.689	Ag2Ba1Sn2	25332	0.668	Sb1	52227	0.801
8	C1Ge1	182363	0.802	Na1Sb1Zn1	645023	0.681	N2Os1	185513	0.666	P1	169539	0.8
9	Si1Sn1	184676	0.786	B1Sb1	184571	0.679	Cd3In1	109285	0.657	In1Mg1	51972	0.799
10	P2Ta1	648187	0.782	Ba1Bi2Zn1	58638	0.679	In1Sb1	659843	0.647	In1	53091	0.798
11	Ge1P1	637492	0.773	Ag1Pb1S3Sb1	24257	0.677	La3Zn22	642095	0.645	In1	57392	0.797
12	Na1O5P1V1	33944	0.766	Ge1Na2Zn1	240728	0.675	Hf1	426944	0.644	Co1Er1Si2	622852	0.796
13	Cd2O7Os2	155761	0.757	Pb1Se2	174577	0.665	Lu1Pb2	104811	0.641	Cu2Eu1Si2	627287	0.796
14	Bi1Te1	617181	0.752	As4Ba3Cd2	424761	0.653	Ag1Er1	58234	0.631	Co1Si2Y1	625129	0.796
15	Ga1Sb1	635312	0.75	As1Br3Ca3	426	0.647	H2Nb1	164606	0.629	Ca5P6Pd6	79096	0.786
16	Ba1Dy2Ni1O5	85046	0.75	Br5C2Ce4	418408	0.644	Al2Ca1Zn2	57550	0.628	Ho1Ni1Si2	639505	0.785
17	As2Mn2O7	69003	0.75	Ca2Sb1	154	0.644	La1Mg3	657966	0.625	Ba1Ga1Ge1	615870	0.784
18	Cr2Hg1Se4	626182	0.747	Al4Na4P12Sr8	409319	0.637	Gd1Zn5	104150	0.624	Eu1Ge1Zn1	246865	0.78
19	B2Be1C2	418618	0.747	Al2Ge2Sr1	608014	0.632	Sn1	52487	0.617	Mg5Tl2	150631	0.777
20	Sn1Te1Zr1	80190	0.746	C7Lu4	83382	0.628	Fe1Sb1Zn1	90397	0.613	Dy1Ni1Si2	658585	0.775

TABLE 4 Top 20 similar materials according to our similarity score to Ba1O3Ti1 (ICSD 183932) and Co1Sb3 (ICSD 164980) over negative and positive energy regions.

	Ba1O3Ti1 (ICSD 183932)						Co1Sb3 (ICSD 164980)					
	Negative			Positive			Negative			Positive		
Rank	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score
1	Ba1O3Ti1	100804	0.989	Ba1O3Ti1	100804	0.936	Fe4P12Th1	200827	0.538	B1H4Li1	95208	0.392
2	Ba1O3Ti1	186462	0.974	Ba1O3Ti1	73639	0.833	As3Co1	610045	0.458	P2Pd3S8	35361	0.376
3	Ba1O3Ti1	73639	0.886	O3Sr1Tc1	183451	0.766	As12Os4Th1	611145	0.443	Ba1Ga2P2	380479	0.357
4	Ba1Mg0.333O3Ta0.667	95495	0.667	B1F3	24783	0.752	Ga2Os1	103785	0.436	F1Lu1O3Se1	417449	0.349
5	Ba3Mg1O9Ta2	240279	0.655	Ba1O3Ti1	186462	0.742	As3Ir1	610737	0.425	Er6I10Ni1	424429	0.344
6	Hf1O3Sr1	89386	0.611	Br1Ce3S8Si2	88691	0.714	Ga2Ru1	635228	0.424	O6Se2Ti1	200203	0.339
7	Ba1Mg0.333Nb0.667O3	95406	0.583	Ag1Hg2O4P1	2208	0.713	Ni3Ta1	105390	0.405	O9Re2V1	92317	0.338
8	Ba3Mg1Nb2O9	240277	0.579	F1O3P1Sn1	2039	0.664	In2Ni3S2	640135	0.404	Pt3Rb2S4	26267	0.335
9	Hf1O3Sr1	161594	0.573	H4Hg2O9P2	413085	0.654	H1Ho1Se1	78957	0.403	Cl6Hg3Te2U1	419437	0.332
10	Bi1In1S3	290195	0.571	O7Sr3Ti2	20294	0.654	La1S2	641808	0.402	B1H4Na1	165835	0.326
11	Br1Cl1	424850	0.559	C1O2	188891	0.651	Ga3Ta1	103976	0.401	Ge4Se10Tl4	26415	0.324
12	Be4N4Sr2	413356	0.549	Bi1Cl2Cu1S1	413289	0.645	In2P2Sr1	260563	0.4	Al2Ru1	609234	0.318
13	Bi2Se3	171571	0.544	C1O2	188893	0.635	H1Se1Y1	72008	0.399	Ba1C1Cl1N1S1	94400	0.312
14	As2Te3	54097	0.526	Br2Ge1	100088	0.635	Nb1Ni3	105175	0.398	Ca2Fe1O6W1	81204	0.308
15	Bi1I1Te1	74501	0.522	Mo1O3	80577	0.634	Ho1Te2Tl1	639771	0.396	C2I1K1N2	40370	0.306
16	Na1O3Ta1	280101	0.518	C1F3H5O5S1	2007	0.629	Sb2Zr1	651784	0.391	F1O3Se1Y1	418898	0.303
17	Al1F3	202681	0.518	Cs3Ni1O2	424578	0.627	Te2Tl1Y1	653098	0.391	Mn1O5Se2	73936	0.303
18	I5In1Sn2	151996	0.515	Ag1Bi1Cl2S1	413290	0.626	Ni3V1	105443	0.391	P1Ru1S1	648023	0.302
19	Mg1Te1	642883	0.505	O3Sr1Ti1	65089	0.615	Hf1Ni3	2415	0.39	Mn1Mo1O4	15615	0.3
20	Ba3Ni1O9Ta2	240281	0.498	Hg2Mo1O4	90084	0.609	Co2Hf1Si2	623793	0.39	Cl2O6Pb1	40286	0.299

Tables 3 and 4 report top 20 similar materials, according to the weighted Pearson coefficient similarity score, for each of the aforementioned four materials over negative and positive energy regions. A close inspection of the density

of states of these materials leads to the following observations.

i. The behavior of the DOS of all these materials is indeed very similar to that of the prototype in both the negative and positive energy region, mostly close to the top of the valence or bottom of conduction band (this is indeed enforced by the choice of the weighted Pearson metric).

ii. The set contains entries that indeed belong to the same class of materials of the prototype (CSn, CGe, etc.); however,

iii. A close look at the overall electronic properties of the materials set shows that the behavior of the DOS is not sufficient to properly predict the behavior of a given system in an actual application.

As an example of Observation *iii*, although Si and Cr5Cs1S8 do show a high degree of similarity in the top of the valence bands in terms of the shape of the DOS, they are very dissimilar materials otherwise. This indicates that a predictive descriptor of materials properties based on electronic structure has to take into account the topological characteristics of the band structure beyond the "mean field" picture of the density of states. Observation *ii* on the other hand, indicates that if one restricts the search space to systems that have, for instance, the same geometry, then the classification becomes extremely accurate and the systems share basically the same (or similar) physical properties. Table 5 illustrates this point by providing the list of 10 materials most similar (according to our similarity score) to Si (ICSD 150530) and Al (ICSD 43492) over negative and positive energy regions with the same number of atoms per cell and geometry. Table 6 gives a similar list for Ba1O3Ti1 (ICSD 183932) and Co1Sb3 (ICSD 164980).

TABLE 5 Top 10 similar materials according to our similarity score to Si (ICSD 150530) and Al (ICSD 43492) over negative and positive energy regions with the same number of atoms per cell and geometry (FCC for both materials).

	Si (ICSD 150530)						Al (ICSD 43492)					
	Negative			Positive			Negative			Positive		
Rank	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score
1	C1Sn1	182365	0.893	B1Sb1	184571	0.679	Ta1	41520	0.482	Tl1	104199	0.64
2	C1Ge1	182363	0.802	As1Sc1	44057	0.589	Nb1	41512	0.461	Ne1	65897	0.492
3	Si1Sn1	184676	0.786	B1P1	181291	0.543	Pa1	77862	0.445	C1	28859	0.471
4	Ga1Sb1	635312	0.75	Be1Te1	290008	0.535	Mg1	180453	0.426	Ar1	53814	0.458
5	As1Ga1	184923	0.73	As1B1	181292	0.532	Br1	168177	0.419	Kr1	43726	0.431
6	B1P1	181291	0.706	Sb1Y1	651741	0.532	V1	41504	0.414	U1	181306	0.416
7	Ga1P1	53963	0.679	Bi1Tb1	617162	0.527	Mo1	41513	0.409	Th1	76039	0.408
8	Be1Te1	290008	0.675	Bi1Dy1	58778	0.514	Ti1	168322	0.403	Xe1	426985	0.4
9	B1Sb1	184571	0.641	Bi1Ho1	43545	0.503	U1	181306	0.398	K1	44669	0.392
10	B1Bi1	184569	0.639	P1Sc1	180831	0.495	Th1	76039	0.363	Au1	426925	0.372

TABLE 6 Top 10 similar materials according to our similarity score to Ba1O3Ti1 (ICSD 183932) and Co1Sb3 (ICSD 164980) over negative and positive energy regions with the same number of atoms per cell and geometry (TET for Ba1O3Ti1 and BCC for Co1Sb3).

	Ba1O3Ti1 (ICSD 183932)						Co1Sb3 (ICSD 164980)					
	Negative			Positive			Negative			Positive		
Rank	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score	Material	ICSD	Score
1	Hf1O3Sr1	161594	0.573	Nb1Te4	601217	0.373	As3Co1	610045	0.458	As3Co1	610045	0.177
2	Na1O3Ta1	280101	0.518	O3Pb1V1	152278	0.337	As3Ir1	610737	0.425	Ni2Zn11	105475	0.094
3	Bi2O3	168807	0.462	O3Pb1Ti1	61169	0.278	Ir1Sb3	640958	0.389	Ir1P3	23713	0.086
4	Na1Nb1O3	280100	0.417	Ga2Mg1Sc2	260213	0.269	Rh1Sb3	650248	0.346	P3Rh1	43724	0.066
5	F4Zr1	35100	0.276	Li1Nd2Si2	642206	0.268	Ni2Zn11	105475	0.316	As3Rh1	611268	0.045
6	Bi2Se3	617096	0.272	Co2P2U1	67932	0.244	As3Rh1	611268	0.278	Rh1Sb3	650248	0.023
7	K1Nb1O3	9535	0.205	C2Mo1Nd2	417666	0.24	Ir1P3	23713	0.221	As3Ir1	610737	0.019
8	I3O1W1	65183	0.195	Si2W3	652552	0.229	O3Re1	55465	0.214	Al12Mo1	58003	0.018
9	Be2Nb3	58722	0.19	Cd1Nd2Ni2	414597	0.226	Co1P3	92393	0.192	Si1V3	52472	0.018
10	B2Ta3	107320	0.189	B2Mo3	614800	0.217	P3Rh1	43724	0.124	Ir1Sb3	640958	0.015

The materials in the neighborhood of a given material (e.g., Si) are all similar to Si, but they may not be similar to each other. To find the largest group of materials in the neighborhood of Si that all share pairwise similarities, we compute a maximum clique in the subgraph induced by the neighborhood of Si. The maximum clique in the network of 20 neighbors of Si with 0.75 cutoff is the following set of 5 materials:

- C1Sn1 (ICSD 182365) - metal,
- Cr5Cs1S8 (ICSD 2566) - insulator,
- Cr3S5Tl1 (ICSD 23632) - insulator,
- Cr5Se8Tl1 (ICSD 37123), and
- P2Ta1 (ICSD 648187) - metal.

The DOS functions are very similar for all these materials, although they differ largely in terms of the band gap.

The maximum clique in the neighborhood of Si with 0.7 cutoff with the same geometry (FCC) and number of atoms per cell (two) consists of the following four materials:

- As1Ga1 (ICSD 184923)
- C1Sn1 (ICSD 182365)
- Ga1Sb1 (ICSD 635312)

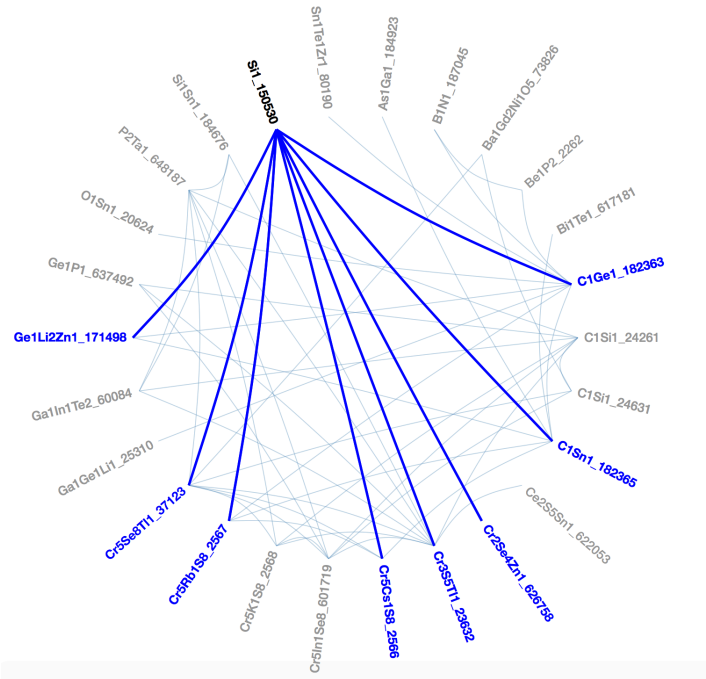


FIGURE 11 Two-hop neighbourhood of Si in a prototype AFLOW library navigation tool based on the developed network representation.

- Si1Sn1 (ICSD 184676)

All these belong to the larger family of FCC semiconductors and indeed share same physical characteristics overall.

3.4 | Network-based Navigation Tool

Finally, Figure 11 provides an illustration of a prototype of a navigation system for AFLOW library based on the network representation developed in this paper. The interactive visualization was implemented as HTML document using JavaScript D^3 library [22] and allows to conveniently navigate the materials in the data base in a web browser. In this illustration, the nodes of the network are limited to the two-hop neighborhood of Si. Pointing the cursor to any node in the network highlights its incident edges that connect it to other nodes. Clicking on an any node in the network will open a web page in AFLOW library of the corresponding material for further inspection. We believe that such a

natural and intuitive navigation tool will significantly enhance users' ability to explore the complex materials data base by visualizing the similarity relationships between the materials using the proposed network representation.

4 | CONCLUSION

This paper develops a methodological framework for construction and analysis of materials maps, based on network representations of a materials database. The proposed network-based approach may provide a valuable tool not only for visualization and navigation of large materials database, but also for carrying out application-specific tasks, such as determining groups of candidate materials for substituting a given material used in a manufacturing process. The work reported in this paper can be viewed as the first step in this direction, opening up many possibilities for future investigations, including the following. (i) Further refining the proposed similarity measures, as well as developing alternative quantification of pairwise relations between materials that could be used to build a network of materials. (ii) Exploring various combinatorial objects representing groups of similar materials in the constructed unweighted materials networks as well as their edge-weighted counterparts, with the edge weights given by the computed similarity scores. Edge-weighted cliques [23, 24, 25] and clique relaxation models [26, 27] are of particular interest in this regard. (iii) Utilizing the proposed methodology in a context of specific applications, such as determining suitable materials for manufacturing composites with desired properties.

REFERENCES

[1] Curtarolo S, Hart GLW, Buongiorno Nardelli M, Mingo N, Sanvito S, Levy O. The high-throughput highway to computational materials design. Nature Materials 2013 Mar;12(3):191–201.

[2] Alberi et al K. The 2019 materials by design roadmap. J Phys D: Appl Phys 2018;52:013001.

[3] Hill J, Mulholland G, Persson K, Seshadri R, Wolverton C, Meredig B. Materials science with large-scale data and informatics: unlocking new opportunities. MRS Bulletin 2016;41:399–409.

[4] Jain A, Hautier G, Ong SP, Persson K. New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships. Journal of Materials Research 2016;31(08):977–994.

[5] Yosipof A, Shimanovich K, Senderowitz H. Materials Informatics: Statistical Modeling in Material Science. Molecular Informatics 2016;35:568–579.

[6] Meredig B. Industrial materials informatics: Analyzing large-scale data to solve applied problems in R&D, manufacturing, and supply chain. Current Opinion in Solid State and Materials Science 2017;DOI: 10.1016/j.cossms.2017.01.003.

[7] Zinoviev D. Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret. Pragmatic Bookshelf; 2018.

[8] Isayev O, Fourches D, Muratov EN, Oses C, Rasch K, Tropsha A, et al. Materials cartography: Representing and mining materials space using structural and electronic fingerprints. Chemistry of Materials 2015;27(3):735–743.

[9] Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor RH, et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. Computational Materials Science 2012;58:227–235.

[10] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 2015;7:20.

[11] Luce RD, Perry AD. A method of matrix analysis of group structure. Psychometrika 1949;14(2):95–116.

[12] Bomze IM, Budinich M, Pardalos PM, Pelillo M. The Maximum Clique Problem. In: Handbook of Combinatorial Optimization, vol. 4 Kluwer Academic Publishers; 1999. p. 1–74.

[13] Wasserman S, Faust K. Social Network Analysis. New York: Cambridge University Press; 1994.

[14] Mokken RJ. Cliques, clubs and clans. Quality and Quantity 1979;13(2):161–173.

[15] Luce RD. Connectivity and generalized cliques in sociometric group structure. Psychometrika 1950;15(2):169–190.

[16] Harary F, Ross IC. A procedure for clique detection using the group matrix. Sociometry 1957;20:205–215.

[17] Cha SH. Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences 2007;1(4):300–307.

[18] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

[19] Barabási AL. Network science. Cambridge University Press; 2016.

[20] Watts DJ, Strogatz SH. Collective dynamics of "small-world" networks. Nature 1998;393(6684):440–442.

[21] Newman MEJ. The structure and function of complex networks. SIAM Review 2003;45:167–256.

[22] Bostock M, Ogievetsky V, Heer J. D^3 data-driven documents. IEEE transactions on visualization and computer graphics 2011;17(12):2301–2309.

[23] Hosseinian S, Fontes DBMM, Butenko S, Buongiorno Nardelli M, Fornari M, Curtarolo S. The Maximum Edge Weight Clique Problem: Formulations and Solution approaches. In: Butenko S, Pardalos PM, Shylo V, editors. Optimization Methods and Applications Springer; 2017.p. 217–237.

[24] Hosseinian S, Fontes DBMM, Butenko S. A nonconvex quadratic optimization approach to the maximum edge weight clique problem. Journal of Global Optimization 2018;72:219–240.

[25] Hosseinian S, Fontes DBMM, Butenko S. A Lagrangian Bound on the Clique Number and an Exact Algorithm for the Maximum Edge Weight Clique Problem. INFORMS Journal on Computing 2020;DOI: 10.1287/ijoc.2019.0898.

[26] Ertem Z, Lykhovyd E, Wang Y, Butenko S. The maximum independent union of cliques problem: complexity and exact approaches. Journal of Global Optimization 2020;76:545–562.

[27] Pattillo J, Youssef N, Butenko S. On clique relaxation models in network analysis. European Journal of Operational Research 2013;226(1):9–18.