Comparing large graphs efficiently by margins of feature vectors

Matthias Dehmer a,b,*,1, Frank Emmert-Streib c,1

a Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Dr. Bohr Gasse 9, A-1030 Vienna, Austria
b University of Vienna, Medical University of Vienna and University of Veterinary Medicine Vienna, Austria
c Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA

Abstract

Measuring the structural similarity of graphs is a challenging and outstanding problem. Most of the classical approaches of the so-called exact graph matching methods are based on graph or subgraph isomorphic relations of the underlying graphs. In contrast to these methods in this paper we introduce a novel approach to measure the structural similarity of directed and undirected graphs that is mainly based on margins of feature vectors representing graphs. We introduce novel graph similarity and dissimilarity measures, provide some properties and analyze their algorithmic complexity. We find that the computational complexity of our measures is polynomial in the graph size and, hence, significantly better than classical methods from, e.g. exact graph matching which are NP-complete. Numerically, we provide some examples of our measure and compare the results with the well-known graph edit distance.

Keywords: Graph similarity; Directed and undirected graphs; Degree vectors; Similarity measures

1. Introduction

Approaches to measure the similarity between graphs have been frequently investigated. This topic can be thematically divided into two main categories: exact and inexact graph matching [2–4,7,12]. Generally, the concept of graph similarity is not uniquely defined. If we consider two graphs, we could refer to several similarity aspects, e.g., semantic similarity, structural similarity, or functional similarity. In this paper, we are interested in methods to measure the structural similarity of graphs. A problem is, that different graph similarity methods emphasize different structural aspects to be measured.

Classical methods to measure the similarity of graphs are based on isomorphic relations. Well-known contributions are from Zelinka [25], Sobik [19,20], and Kaden [13,14]. For example, the Zelinka-distance is based on the principle that two graphs are more similar, the bigger the common induced subgraph is [25].
means, that graphs which have a large common induced subgraph have a small distance and vice versa [25,19]. Zelinka introduced such a measure based on isomorphic relations for unlabeled graphs. Then, Sobik [19,20] and Kaden [13,14] generalized this measure for arbitrary and also labeled directed and undirected graphs. For practical use and for processing graph corpora of large graphs, the complexity of these approaches is considered to be unacceptable because the underlying problems are mostly NP-complete [22,24].

In contrast to the methods which are based on isomorphic relations there exists an important class of graph similarity measures based on the well-known graph edit distance (inexact graph matching) originally due to Bunke [2–4]. The graph edit distance is based on basic weighted transformation steps, i.e., deletions, substitutions, and insertions of vertices and edges. More precisely, Bunke [2–4] denotes as an optimal inexact match a sequence of transformations which transform a graph \( G^1 \) into \( G^2 \) by producing minimal transformation costs. If we now assume that \( m_1, m_2, \ldots, m_n \) are all possible inexact matches between \( G^1 \) and \( G^2 \), then the optimal inexact match [2] \( m^* \) is defined by \( c(m^*) = \min \{ c(m_i) \mid 1 \leq i \leq n \} \), where \( c(m_i) \) denotes the costs of \( c_i \). The similarity of the graphs is defined as the minimum cost of transformations. Furthermore, Shapiro [18] developed a well-known principle for determining the structural similarity of graphs, based on the adjacency matrices. In order to outline Shapiro’s method shortly, we assume two graphs \( G^1 = (V^1, E^1) \), \( G^2 = (V^2, E^2) \) and let \( (M^1_{ij}), (M^2_{ij}) \), \( 1 \leq i, j \leq |V^1|, 1 \leq i, j \leq |V^2| \) be the corresponding adjacency matrices. Now, by permuting the rows and columns of \( (M^2_{ij}) \), \( (M^1_{ij}) \) such that the matrix elements conform with the matrix elements of \( (M^1_{ij}) \) as much as possible. Hence, Shapiro [18] defines the graph distance between \( G^1 \) and \( G^2 \) by the minimal number of disentangling matrix elements. To finish our review of related work on methods to measure the structural similarity between graphs, we outline a known contribution of Gernert [8]. The contribution of Gernert [8] describes the task of determining the structural similarity between graphs based on graph grammars [9,10]. In order to provide the main result of Gernert [8], we consider a set of graphs \( S := \{ G^1, G^2, \ldots, G^p \} \), where \( G^i, 1 \leq i \leq n \) are assumed to be connected. Furthermore, a graph grammar \( z \) is assumed, that generates a graph set \( \bar{S} := \{ \bar{G}^1, \bar{G}^2, \ldots, \bar{G}^p \}, S \subseteq \bar{S} \). Starting from

\[
f(\bar{G}^i, \bar{G}^k) := \begin{cases} 0 & : \bar{G}^i \text{ isomorphic to } \bar{G}^k \\ 1 & : \bar{G}^i \rightarrow \bar{G}^k \text{ one replacement step} \\ \text{undefined} & : \text{else} \end{cases}
\]

Gernert defines the so-called path lengths between \( \bar{G}^i \) and \( \bar{G}^k \), if \( \bar{G}^i \) can be generated by a finite sequence of steps, starting from \( \bar{G}^i \). Due to the fact that it exists at least one graph, e.g., the start graph, \( G^* \) with \( G^* \rightarrow \bar{G}^i \), \( G^* \rightarrow \bar{G}^k \) and \( l \) denotes the path length Gernert [8] obtains

\[
d(\bar{G}^i, \bar{G}^k) := \min \{ l(G^*, G^i) + l(G^*, G^k) \mid G^* \in \bar{S} \land G^* \subseteq G^i \land G^* \subseteq G^k \}.
\]

A thorough discussion of existing graph similarity approaches can be found in, e.g. [4,7,14,17].

In contrast to these methods we define in this paper a fast method for determining the structural similarity between directed and undirected graphs based on feature vectors. The main idea of the new method is based on the so-called margins that are obtained as differences of feature vectors representing the graphs. In this paper, we use the out- and in-degree vectors for directed and just the degree vectors for undirected graphs as feature vectors. One main advantage of our novel principle is that degree sequences can be easily determined with computational complexity \( O(|V|^2) \) from parsing the corresponding adjacency matrices. We want to notice that in our case the process of inferring the margins from the given degree vectors of corresponding graphs has been inspired by the theory of vector majorization. Generally, majorization [16,23] was first developed by Lorenz [15] for comparing income distributions by using inequalities.

This paper is organized as follows. In Section 2, we introduce a novel similarity and dissimilarity measure to determine the structural similarity between directed and indirected graphs. Section 3 briefly outlines the relatedness to some other classical graph similarity methods. In Section 4, we analyze the time complexity of the novel graph similarity and dissimilarity measures. In Section 5, we compare our approach to measure the structural similarity with the graph edit distance numerically. The paper finishes in Section 6 with a summary and conclusion.
2. Graph similarity measures based on margins of feature vectors

We start with some definitions we need for our further studies.
Let $G = (V, E), E \subseteq V \times V, |V| < \infty$ be a finite and directed graph. Then we define the following entities:

\begin{align*}
\mathcal{N}_+ (v) := & \{ \tilde{v} \in V \setminus \{v\} | (v, \tilde{v}) \in E\}, \\
\mathcal{N}_- (v) := & \{ u \in V \setminus \{v\} | (u, v) \in E\}, \\
\delta_{\text{out}}(v) := & |\mathcal{N}_+ (v)|, \\
\delta_{\text{in}}(v) := & |\mathcal{N}_- (v)|.
\end{align*}

(1) (2) (3) (4)

The set $\mathcal{N}_+ (v)$ contains all vertices of the graph $G$ that are connected with the node $v$ by an edge and $\delta_{\text{out}}(v)$ is just the number of elements in that list corresponding to the so-called out-degree of node $v$. Analogously, the in-degrees are defined. This gives us the degree vectors $(\delta_{\text{out}}(v_1), \delta_{\text{out}}(v_2), \ldots, \delta_{\text{out}}(v_{|V|}))$ and $(\delta_{\text{in}}(v_1), \delta_{\text{in}}(v_2), \ldots, \delta_{\text{in}}(v_{|V|}))$ of all out- and in-degrees for graph $G$ ordered by their vertex index. In the following we will need the ordered degree vector in decreasing order resulting in:

\[ s_{\text{out}}(G) := (\delta_{\text{out}}(v_{(1)}), \delta_{\text{out}}(v_{(2)}), \ldots, \delta_{\text{out}}(v_{(|V|)})), \]

(5)

and

\[ s_{\text{in}}(G) := (\delta_{\text{in}}(v_{(1)}), \delta_{\text{in}}(v_{(2)}), \ldots, \delta_{\text{in}}(v_{(|V|)})). \]

(6)

That means, e.g., for the out-degrees it hold always $\delta_{\text{in}}(v_{(j)}) \geq \delta_{\text{in}}(v_{(j+1)}), s_{\text{out}}(G) \in \mathbb{N}^{|V|}$ and $s_{\text{in}}(G) \in \mathbb{N}^{|V|}$ are the permuted out-degree and in-degree vectors of $G$, but for simplicity we just call them out- and in-degree vectors. In the case of undirected graphs we simply speak of degree vectors instead of out-degree and in-degree vectors.

**Definition 2.1.** Let $X$ be a set. A positive real function $s : X \times X \to [0, 1]$ is called similarity measure if

\begin{align*}
s(x, y) = & s(y, x) \quad \forall x, y \in X, \\
s(x, y) \leq & s(x, x) \quad \forall x, y \in X, \\
s(x, x) = & 1 \quad \forall x \in X.
\end{align*}

(7) (8) (9)

**Remark 2.1.** We want to remark that $d(x, y) := 1 - s(x, y) \in [0, 1]$ defines a measure of dissimilarity for $x, y \in X$.

**Lemma 2.2.** Let $x, y, \sigma \in \mathbb{R}$ and let be $s(x, y)$ the positive real function

\[ s(x, y) := e^{-\frac{|x-y|^2}{\sigma^2}}. \]

(10)

$s(x, y)$ is a similarity measure in the sense of **Definition 2.1**.

**Proof.** The symmetry condition $s(x, y) = s(y, x) \forall x, y \in X$ follows immediately from the definition of $s(x, y)$. Further, we obtain that $s(x, y) \in [0, 1]$, because $s(x, x) = e^0 = 1$ and

\[ \frac{1}{e^{-\frac{|x-y|^2}{\sigma^2}}} \leq 1. \]

(11)

That means we also get the property $s(x, y) \leq s(x, x) = 1 \forall x, y \in X$. \qed

The following proposition proves that the additive composition of similarity measure is also a similarity measure. This technique is also essential in machine learning where this fact is often used, e.g., that the sum of the so-called kernel functions is also a kernel function [6].
**Proposition 2.3.** Let \( s_j(x_j, y_j), \ 1 \leq j \leq k, \ k > 1 \) be similarity measures in the sense of **Definition 2.1**. Then,

\[
s(x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_k) := \frac{1}{k} \sum_{j=1}^{k} s_j(x_j, y_j) \quad \forall x_j, \ y_j \in X_j
\]  

(12)

is also a similarity measure.

**Proof.** By assumption it holds \( s_j(x_j, y_j) \in [0, 1] \). By using Eq. (12) we also obtain

\[
s(x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_k) \in [0, 1].
\]  

(13)

Since each similarity measure \( s_j(x_j, y_j) \) is symmetric we conclude that

\[
\frac{1}{k} \sum_{j=1}^{k} s_j(x_j, y_j) = \frac{1}{k} \sum_{j=1}^{k} s_j(y_j, x_j).
\]  

(14)

Finally, from \( s_j(x_j, y_j) \in [0, 1], \ 1 \leq j \leq k, \ k > 1 \), we infer that

\[
\frac{1}{k} (s_1(x_1, y_1) + s_2(x_2, y_2) + \cdots + s_k(x_k, y_k)) \leq 1. \quad \square
\]  

**Remark 2.4.** We want to remark that **Proposition 2.3** holds analogously for measures of dissimilarity,

\[
d(x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_k) := \frac{1}{k} \sum_{j=1}^{k} d_j(x_j, y_j).
\]  

(16)

**Definition 2.2.** Let \( G^1 = (V^1, E^1) \) and \( G^2 = (V^2, E^2) \) be finite and directed graphs without reflexive and multiple edges. Furthermore, let \( n \) be defined as \( n := \max(|V_1|, |V_2|) \). By adding zero vector entries we can always obtain two vectors of equal length \( n \). Starting from

\[
s_{\text{out}}(G^1) = (\delta_{\text{out}}(v^G_{1(1)}), \delta_{\text{out}}(v^G_{1(2)}), \ldots, \delta_{\text{out}}(v^G_{1(n)})),
\]  

(17)

\[
s_{\text{in}}(G^1) = (\delta_{\text{in}}(v^G_{1(1)}), \delta_{\text{in}}(v^G_{1(2)}), \ldots, \delta_{\text{in}}(v^G_{1(n)})),
\]  

(18)

we define the numbers \( x^0_1, x^0_2, \ldots, x^0_n \in \mathbb{N} \) and \( x^1_1, x^1_2, \ldots, x^1_n \in \mathbb{N} \), by

\[
x^0_1 = \| \delta_{\text{out}}(v^G_{1(1)}) - \delta_{\text{out}}(v^G_{1(2)}) \|,
\]  

(19)

\[
x^0_2 = \| \delta_{\text{out}}(v^G_{1(2)}) - \delta_{\text{out}}(v^G_{1(3)}) \|,
\]  

(20)

\[\vdots\]

\[
x^n = \| \delta_{\text{out}}(v^G_{1(n)}) - \delta_{\text{out}}(v^G_{1(n+1)}) \|.
\]  

(21)

and

\[
x^1_1 = \| \delta_{\text{in}}(v^G_{1(1)}) - \delta_{\text{in}}(v^G_{1(2)}) \|,
\]  

(22)

\[
x^1_2 = \| \delta_{\text{in}}(v^G_{1(2)}) - \delta_{\text{in}}(v^G_{1(3)}) \|,
\]  

(23)

\[\vdots\]

\[
x^n = \| \delta_{\text{in}}(v^G_{1(n)}) - \delta_{\text{in}}(v^G_{1(n+1)}) \|.
\]  

(24)

We call the \( x_i \)’s the out- and in-degree margins between the corresponding out- and in-degree values at rank position \( i \).
Based on Definition 2.2 we now define a similarity measure depending on
\[ x_1^{\text{out}}, x_2^{\text{out}}, \ldots, x_n^{\text{out}} \quad \text{and} \quad x_1^{\text{in}}, x_2^{\text{in}}, \ldots, x_n^{\text{in}} \in \mathbb{N}, \] (25)
Because the out- and in-degree margins capture structural information of the underlying graph the resulting measure covers structural information of the underlying graphs too.

Corollary 2.5. Let \( x_1, x_2, \ldots, x_n \) be integer numbers, e.g., the out- or in-degree margins. Then we define
\[ s(x_1, x_2, \ldots, x_n, \sigma) := \frac{e^{-\frac{(x_1)^2}{n}} + e^{-\frac{(x_2)^2}{n}} + \cdots + e^{-\frac{(x_n)^2}{n}}}{n}. \] (26)
\( s(x_1, x_2, \ldots, x_n, \sigma) \) is a similarity measure.

Proof. Assuming \( x_j \in \mathbb{N}, 1 \leq j \leq n, \sigma \in \mathbb{R} \) we obtain with Lemma 2.2 that \( e^{-\frac{(x_j)^2}{n}} \) is a similarity measure. If \( x_1 = x_2 = \cdots = x_n = 0 \), then it holds \( s(x_1, x_1, \ldots, x_1, \sigma) = 1 \). From this we get also the inequality
\[ \frac{e^{-\frac{(x_1)^2}{n}} + e^{-\frac{(x_2)^2}{n}} + \cdots + e^{-\frac{(x_n)^2}{n}}}{n} \leq 1. \] (27)
Hence, we also conclude that \( s(x_1, x_2, \ldots, x_n, \sigma) \leq s(x_1, x_1, \ldots, x_1, \sigma) = 1. \)

Based on these results we are now able to introduce a novel graph similarity measure based on the margins of out- and in-degree vectors. The following theorem shows that the following measured values have the properties of a classical similarity measure (cf. Definition 2.1).

Theorem 2.6. Let \( G^1 = (V^1, E^1) \) and \( G^2 = (V^2, E^2) \) be directed graphs without reflexive and multiple edges. Following Definition 2.2 we obtain the margins \( x_1^{\text{out}}, x_2^{\text{out}}, \ldots, x_n^{\text{out}} \) and \( x_1^{\text{in}}, x_2^{\text{in}}, \ldots, x_n^{\text{in}} \). Based on Corollary 2.5 we define for \( p_1, p_2 \in [0,1] \) and \( p_1 + p_2 = 1 \), the measure
\[ s(G^1, G^2) = p_1 \cdot s(x_1^{\text{out}}, x_2^{\text{out}}, \ldots, x_n^{\text{out}}, \sigma^{\text{out}}) + p_2 \cdot s(x_1^{\text{in}}, x_2^{\text{in}}, \ldots, x_n^{\text{in}}, \sigma^{\text{in}}). \] (28)
Then \( s(G^1, G^2) \) is a graph similarity measure.

Proof. The symmetry condition of the local similarity functions
\[ s(x_1^{\text{out}}, x_2^{\text{out}}, \ldots, x_n^{\text{out}}, \sigma^{\text{out}}) \in [0,1] \quad \text{and} \quad s(x_1^{\text{in}}, x_2^{\text{in}}, \ldots, x_n^{\text{in}}, \sigma^{\text{in}}) \in [0,1] \]
follows immediately from Lemma 2.2 and Corollary 2.5. If the underlying graphs \( G^1 = (V^1, E^1) \) and \( G^2 = (V^2, E^2) \) are identical, then we get with Definition 2.2 that \( x_1^{\text{out}} = x_2^{\text{out}} = \cdots = x_n^{\text{out}} = 0 \) and \( x_1^{\text{in}} = x_2^{\text{in}} = \cdots = x_n^{\text{in}} = 0 \). Applying Corollary 2.5, we conclude that
\[ s(x_1^{\text{out}}, x_2^{\text{out}}, \ldots, x_n^{\text{out}}, \sigma^{\text{out}}) = s(x_1^{\text{in}}, x_2^{\text{in}}, \ldots, x_n^{\text{in}}, \sigma^{\text{in}}) = 1. \]
Then, we also conclude that \( s(G^1, G^1) = p_1 + p_2 = 1 \). The other direction of this assertion is also true. In the case that \( G^1 = (V^1, E^1) \) and \( G^2 = (V^2, E^2) \) are not identical it holds,
\[ s(x_1^{\text{out}}, x_2^{\text{out}}, \ldots, x_n^{\text{out}}, \sigma^{\text{out}}) \leq 1 \quad \text{and} \quad s(x_1^{\text{in}}, x_2^{\text{in}}, \ldots, x_n^{\text{in}}, \sigma^{\text{in}}) \leq 1. \]
Finally, we get
\[ s(G^1, G^2) = p_1 \cdot s(x_1^{\text{out}}, x_2^{\text{out}}, \ldots, x_n^{\text{out}}, \sigma^{\text{out}}) + p_2 \cdot s(x_1^{\text{in}}, x_2^{\text{in}}, \ldots, x_n^{\text{in}}, \sigma^{\text{in}}) \leq p_1 + p_2 = 1, \]
i.e., we have proven that \( s(G^1, G^2) \leq s(G^1, G^1) = 1. \)

Fig. 1 shows the extraction process of the margins from out-degree and in-degree vectors. It is clear, that there is a large number of similarity or dissimilarity measures that can be defined based on the margins of out- and in-degree sequences. We just want to provide one further dissimilarity measure which is very appealing due to its simplicity. We define this measure only for undirected graphs, however, an extension to directed graphs follow immediately.
Theorem 2.7. Let $G^1 = (V^1, E^1)$ and $G^2 = (V^2, E^2)$ be undirected graphs without reflexive and multiple edges.

Let $s(G^1), s(G^2)$ be the degree vectors and $D^2_N$ the highest degree of both degree vectors.

$d(x, y)$ be the positive rational function:

$$d(G^1, G^2) = \frac{1}{n \cdot A} \sum_{i=1}^{n} ||s(G^1)(i) - s(G^2)(i)||.$$  (29)

Then $d(G^1, G^2)$ is a dissimilarity measure that depends only on the margins of the degree vectors, i.e., $d(G^1, G^2) = d(\alpha_1, \alpha_2 \ldots, \alpha_n)$.

Proof. If $G^1, G^2$ are identical, it holds $d(G^1, G^1) = 0$ and vice versa. Further, we get from the definition that $d(G^1, G^2) \in [0, 1]$. □

We call the distance measure based on Eq. (29) the margins of feature vectors, $d_{MFV}$. From this measure we obtain straightforward two conclusions.

Corollary 2.8. Let $G^1$ and $G^2$ be two undirected graphs. Then, the following assertion:

$$d_{GED}(G^1, G^2) = 0 \Rightarrow d_{MFV}(s(G^1), s(G^2)) = 0$$  (30)

holds.

With $d_{GED}$ we denote the already mentioned graph edit distance (cf. Section 1). Fig. 2 gives a counter example that the conclusion in the opposite direction does not hold in general because the degree vectors (out- or in-degrees) do not determine the network topology uniquely. This leads us directly to the next corollary.
Corollary 2.9. If two graphs $G_1$ and $G_2$ have the same degree vectors but a different connectivity then $d_{\text{MFV}}(G_1, G_3) = d_{\text{MFV}}(G_2, G_3)$ holds if compared to an arbitrary graph $G_3$.

That means our measure defined in Eq. (29) is invariant under transformations that leave the degree sequences unchanged but change the connectivity of a graph. We want to remark that this invariance property is not per se bad or good but it depends on the task. If we want to find graphs that are invariant according to this property our measure is appropriate but if we want to distinguish between different graphs with a different structure but the same degree sequence the measure defined so far might not obtain the desired results. However, due to Proposition 2.3 and Remark 2.4 we can easily modify our measure by addition of a second or even more similarity/dissimilarity measures, also based on the comparison of feature vectors, to capture more structural information of the underlying graphs. Here, we just want to give one example of such an additional term that allows to distinguish between a graph and its invariant as defined in Fig. 2. If we map a graph to its line graph $[1,11]$ that means if we map each vertex to an edge and each edge to a vertex and connect vertices in the new graph if both edges in the old graph are connected by the same vertex we obtain the corresponding line graphs in Fig. 3. One can see that the degree sequence of the line graphs are no longer identical but can be distinguished. We define now the composite measure and prove that it is a dissimilarity measure.

Theorem 2.10. Let $G_1$ and $G_2$ be two undirected graphs and $s(G_1), s(G_2)$ the degree sequences. Correspondingly, we have their line graphs $GL_1, GL_2$ and their degree sequences $s(GL_1), s(GL_2)$. We calculate the margins of these degree sequences by

$$a_i := \|s(G_1) - s(G_2)\|_i,$$

$$x_{j}^L := \|s(GL_1) - s(GL_2)\|_j,$$

for $i \in n$ with $n = \max\{|V_1|, |V_2|\}$ and $j \in n_L$ with $n_L = \max\{|V_1^L|, |V_2^L|\}$. From this it follows that

$$d(G_1, G_2) := \frac{p_1}{n \cdot A_1} \sum_{i=1}^{n} a_i + \frac{p_2}{n_L \cdot A_2} \sum_{j=1}^{n_L} x_{j}^L$$

is a dissimilarity measure for $p_1, p_2 \in [0, 1]$ and $p_1 + p_2 = 1$.

Proof. If $G_1, G_2$ are identical, it holds also by application of Proposition 2.3 and Remark 2.4 that $d(G_1, G_2) = p_1 \cdot 0 + p_2 \cdot 0$ and vice versa. In the case that $G_1 \neq G_2$ we get $d(G_1, G_2) \leq p_1 \cdot 1 + p_2 \cdot 1 = 1$. □

3. Relatedness to other graph similarity methods

In Section 2, we introduced novel measures (similarity and dissimilarity measures) for measuring the structural similarity of directed and indrected graphs. In order to distinguish our novel approach from some classical graph similarity methods, we now briefly outline their main construction characteristics and compare...
these characteristics with those of our novel method. As we have already mentioned in Section 1, most of the classical methods for measuring the structural similarity methods are from the so-called exact graph matching [4]. That means that the resulting graph similarity measures are mainly based on graph and subgraph isomorphic relations. Then, measuring the structural similarity between two graphs is based on the principle of finding a maximal common induced and isomorphic subgraph [13,14,19,20,25]. In Fig. 4, we depict the procedure of finding a maximal common induced isomorphic subgraph for two given labeled and directed graphs. This basically explains the task of exact graph matching: finding an induced subgraph pattern in the first graph which exactly matches another graph pattern in the second graph. Based on the determined common induced isomorphic subgraph, graph similarity measures can be obtained by, e.g., the Zelinka-distance [25]. The main difference compared to our novel approach for measuring the structural similarity of directed and undirected graphs is that we obtain the structural information which indicates the similarity between the given graphs by a quasi inexact comparison of the margins using local similarity or dissimilarity measures. We notice that the degree vectors cover structural information of the underlying graphs. Finally, our novel graph similarity approach belongs not to the exact graph matching paradigm but to the inexact paradigm using a feature vector approach. A different approach, also based on feature vectors, is due to Champin et al. [5]. The approach of Champin et al. builds on the derivation of feature vectors, e.g., by using vertex and edge labels and then applies the well known similarity concept of Tversky [21] to the obtained feature vectors.

We furthermore notice that there is a significant difference between the classical graph edit distance by Bunke [2–4], from the inexact graph matching paradigm, and our method. The GED approach mainly deals with defining weighted graph transformation steps, i.e., deletions, substitutions, and insertions of vertices and edges. Then, the similarity between two given graphs is related with the sequence of graph transformations producing minimal transformation costs to map one graph onto the other. But in our case we extract the structural information for finding the similarity and dissimilarity measure from the corresponding margins of degree vectors. Based on the inferred margins from degree vectors we inexactly compare these feature vectors by using our novel similarity and dissimilarity measures. This approach also belongs to the inexact graph matching paradigm. We further notice that the sequence of graph transformations producing minimal transformation costs (GED) is theoretically not directly comparable with our measures based on the extracted margins.

4. Complexity of our measure

In this section, we analyze the computational complexity of our novel approach to measure the structural similarity of graphs. Our main assumption is that after deriving the out-degree and in-degree vectors (in the directed case, otherwise we only deal with simple degree vectors) that the components are given in a decreasing order. Then the following assertion can be made.

Lemma 4.1. Let \((s_1, s_2, \ldots, s_n)\) be a sequence of natural numbers. The time complexity of an algorithm for obtaining an ordering in such a way that \(s_1 \geq s_2 \geq \cdots \geq s_n\) is at most \(O(n^2)\).
Corollary 4.2. Let \( s(G) \) be degree vector of a non-empty graph \( G = (V, E), |V| = n \). Creating a decreasing order of the components of \( s(G) \) requires time complexity \( O(n^2) \).

Proposition 4.3. The computation of the Similarity Measure (28) (Theorem 2.6) and Dissimilarity Measure (29) (Lemma 2.7) requires time complexity \( O(n^2) \).

Proof. Applying Lemma 4.4, we get that the creation of a decreasing ordering of degree vector components of a graph requires time complexity \( O(n^2) \). But the computation of both measures requires time complexity \( O(n) \). By adding up the two complexities we obtain that the resulting time complexity is also \( O(n^2) \). □

Corollary 4.4. Proposition 4.3 also holds for directed graphs using out-degree and in-degree vectors.

Lemma 4.5. Let \( G = (V, E), |V| = n \) be a non-empty undirected graph. The line graph is formally given by \( G'^{\ell} = (E, E'^{\ell}) \), where
\[
E'^{\ell} := \{ (e, e') | e, e' \in E \text{ and } e, e' \text{ are incident in } G \}.
\]
Then, the computation of \( G'^{\ell} \) requires time complexity \( O(n^2) \).

Proof. In order to map the edges of \( G \) to vertices of \( G'^{\ell} \), we must parse the whole adjacency matrix which have \( n^2 \) entries. The remaining operations require constant time complexity, i.e., \( O(1) \). □

Proposition 4.6. Let \( G = (V, E), |V| = n \) be a non-empty undirected graph and let \( G'^{\ell} = (E, E'^{\ell}) \) be this line graph. The algorithmic computation of graph dissimilarity measure based on Eq. (33) requires time complexity \( O(n^3) \).

Proof. From Lemma 4.5 we know that the line graph computation has time complexity \( O(n^2) \). The computation of Eq. (33) needs altogether computational complexity \( O(n) \). □

5. Numerical results

In this section, we apply our newly introduced graph dissimilarity measure based on Theorem 2.10 to a set of graphs and discuss the obtained numerical results to gain a better understanding about its performance. Also, we apply the graph edit distance [2] to the same set of graphs. From the comparison of the results of both graph measures we learn how like or alike both measures are.

Because we want to apply our measure (Eq. (2.10)) and the graph edit distance to the same set of graphs \( GS \) we need to choose \( |V| \) sufficiently small to enable that also the GED can be applied. For this reason, we choose \( |V| = 8 \). For simplicity we restrict our numerical analysis to undirected graph. The set of graphs is now generated in the following way. We enumerate the edges of a graph with \( |V| \) nodes by
\[
\begin{pmatrix}
0 & 1 & 2 & \ldots & (|V| - 1) \\
1 & 0 & |V| & \ldots & (2|V| - 3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
(|V| - 1) & (2|V| - 3) & (3|V| - 6) & \ldots & (|V|^2 - |V|)/2
\end{pmatrix}
\]
(34)
The graph \( G_i \) contains \( i \) edges, i.e, all edges up to number \( i \). By this procedure we generate \( (|V|^2 - |V|)/2 = 28 \) graphs.

Numerical results for this study are shown in Fig. 5. The left figure shows the values of \( d_{G_{ED}} \) plotted against \( d_{MVF} \) for all possible graph pairs \( (G_i, G_j) \) for \( i \neq j \) and \( G_i, G_j \in GS \). The overall mean error from this comparison is 11.7%. That means, in average both dissimilarity measures differ by 11.7%. To evaluate this value we generate randomly values in \([0, 1]\), calculate the difference and average over the ensemble of these values. From this we obtain a mean difference of 33%. That means, our measure approximates the GED surprisingly well considering the fact, that we did not design our measure to aim to approximate the GED as good as possible.
In the right Fig. 5, we compare the rank position of both measures instead of the values. The overall rank error is 12.6% which corresponds to a shift of 51 places in average. A randomization of rank positions gives a shift of 136 places in average (the total number of graph pairs is 406).

Now, we want to take a closer look to the obtained results by presenting same examples of graph pairs that resulted in similar or dissimilar dissimilarity values. In Fig. 6, we show two examples of graph pairs that have been detected as similar (low dissimilarity value) by both measures and in Fig. 7 we show one graph pair that has been detected as unsimilar (high dissimilarity value) by both methods. Visual inspection of these graph pairs confirms these results intuitively. Fig. 8 shows a more controversial example of graph pairs. Numerical

Fig. 5. We calculated the $d_{GED}$ and $d_{MFV}$ distances for all pairs of different graphs from the set $GS$ which consists of 29 graphs. This gives a total of 406 different graph pairs. (Left) Shown are the absolute values of the distance measures by plotting $d_{GED}(G_i, G_j)$ against $d_{MFV}(G_i, G_j)$ for $i \neq j$ and $G_i, G_j \in GS$. The overall error is 11.7%. (Right) The rank of graph pair $(G_i, G_j)$ from $d_{GED}$ is plotted against the rank of these pair from $d_{MFV}$. The overall rank error is 12.6% (51 places).

![Graphs](image)

Fig. 6. Some examples of graph pairs from the set $GS$ that are detected as similar by both methods. (Top) $G_2$ and $G_4$, $d_{GED}(G_2, G_4) = 0.07$, $d_{MFV}(G_2, G_4) = 0.07$. (Bottom) $G_{18}$ and $G_{20}$, $d_{GED}(G_{18}, G_{20}) = 0.07$, $d_{MFV}(G_{18}, G_{20}) = 0.16$. 
calculation gives $d_{\text{GED}}(G_{18}, G_{28}) = 0.36$ and $d_{\text{MFV}}(G_{18}, G_{28}) = 0.71$. In this case, the result of both measures is significantly different. The interesting observation regarding this example is that we can find arguments supporting the similarity or dissimilarity of this graph pair, hence, even in cases in which both methods produce different results the outcome is meaningful if one chooses the right argument.

6. Summary and conclusions

In the present paper, we introduced a very efficient method to determine the structural similarity of directed and undirected graphs. Our novel approach is mainly based on so called margins, which are obtained as differences of degree vectors, covering the structural information of the given graphs. We defined measures for directed and undirected graphs and proved that these measures are similarity and dissimilarity measures. Furthermore, we examined the computational complexity of these measures and showed that they can be calculated in $O(n^2)$ time steps, whereas $n$ corresponds to the number of vertices in the larger graph. This polynomial dependency on the systems (graph) size enables an application to large graphs. This is in contrast to classical methods, e.g., subgraph isomorphism or the graph edit distance, that are proven to be NP-complete [22,24]. Interestingly, numerical results for a test set of graphs showed that the numerical differences between our measure and the graph edit distance [2–4] is relatively small. This is an important observation and further studies will be conducted to investigate this relation in more detail.

Acknowledgement

Financial support from the Vienna Science and Technology fond (WWTF) to Matthias Dehmer and Arndt von Haeseler is gratefully acknowledged.
References